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*—That is why this work is dedicated to considering the model of data processing that is based on cloud and volunteer computing for big data processing. Through adoption of the strengths that are inherent with the two environments, the model seeks to attain optimum utilisation of resources, improvement in performance and decrease in costs. The function of HR Alloc contributes to the dynamic data placement and resource distribution: the results obtained show an increase in processing speed and the optimization of costs compared to cloud-based solutions.**

***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:

Innovative competence acquisition is a strategic management process by which organisations obtain high levels of innovative competence as a means of achieving competitive advantage in their respective industries. This industry-based view of competitive advantage led me to the following research statement and conjecture. The presented research objectives shall therefore to propose a novel hybrid data processing architecture involving both cloud and volunteer computing environments with an intention of dealing with a big data processing's main problem that are resources and cost. Existing big data architectures mostly employ cloud compute as their single big data processing method which although provides an efficient solution with scalability and reliability, can become very costly when it comes to sustained and enormous scale data processing. On the other hand, volunteer computing relies on the surplus processing power of computers owned by people proving itself to be a cheap resource but is highly unreliable as compared to cloud services. Therefore, based on the synergism that is expected to take place from a combination of the cloud with the volunteer computing environment, we expect our approach to reap enhanced performance and substantial cost reduction. That is the concept of how the job responsibilities can be intelligent and balanced according to the available resources with regards to cost parameters. Resource-intensive procedures and tasks that require high reliability and speed can be submitted to the cloud, while less crucial or the ones that can be executed in parallel in a volunteer network. The following major strategic allocation has been proposed with the objective of bringing down the utilization of costly cloud-based resources and at the same time, minimizing expenditure while delivering results of equal quality.

This approach is based on the newly proposed HR Alloc algorithm which is used to manage the placement of the data and the resources in such hybrid infrastructures. We presume that by applying this algorithm, our model will increase the usage of the available resources and improve the efficiency and effectiveness of big data analysis. This hypothesis is derived on the assumption that a synergy of the characteristics of cloud and volunteer computing platforms would result in better scalability, flexibility and cost optimization compared with mere cloud solutions. Specifically, we anticipate that our hybrid model will demonstrate significant improvements in the following areas:

A. Performance:

Hence, by utilizing the principal computational capacity of cloud computing for such workloads requiring high performance, Moreover, by incorporating volunteer resources for other workloads, we anticipate that our model would cut down processing times and improve system performance.

B. Cost Efficiency:

According to the hybrid model, the utilization of cloud resources is likely to be lowered in an aim to minimize operational costs. Volunteer computing represents one of the largest resources of the pool, which makes virtually eliminates the costs related to data processing tasks.

C. Scalability:

Scalability: Extending volunteer computing with resources from a cloud computing infrastructure is believed to improve

the scalability of the system based on resource requirements that may be needed to process bigger data sets with higher data variability.

*D. **Resource Optimization:***

Flexible assignment of tasks associated with the real-time availability of computer resources will help achieve an approximate distribution of tasks across the available computational capacities, which will prevent the overloading of specific resources.

II. **Related Work (Literature Survey):**

In preprocessing our research, we applied a scholarly review of literature in areas concerning big data analytics, cloud computing, and volunteer computing. In this review, recent publications and important research on similar integrations and technology enhancements in these fields were also involved. 1. The study by Dos Anjos et al is one of the most widely cited studies and serves as a starting point for our research proposal. Their work, "Data Processing Model to Perform Big Data Analytics in Hybrid Infrastructures", discusses the integration of computing infrastructures such as cloud and volunteer for big data analytics. To summarize, the authors discuss in detail the prospects and problems of using hybrid infrastructures and develop a model of data processing, which will allow to maximize the use of resources and achieve high performance indicators. Based on this, their findings show that they can realize relatively large gains in cost efficiency and in processing capacity when adopting hybrid models of decision-making, which is compatible with our stated goals of this undertaking. 2. Marz and Warren (2015): In their book "Big Data: Fundamentals of Scalable Realtime Data Systems," Marz and Warren provide a clearer understanding of the principles as well as the best practices when it comes to creating large-scale, massive real-time data systems. Their work focuses support on the aspects of scaling, managing faults, and processing streams on big data in real-time. The approaches mentioned in their book can be considered as the theoretical background for the hybrid model in the view of the opportunities for achieving scalability and reliability of the data processing tasks. About processing of the data in the real-time and batch modes their focus on the Lambda Architecture played a crucial role, so we followed the same strategy to deal with different data processing needs. 3. Zaharia et al. (2010): The study by Zaharia et al. , titled "Spark: In the paper titled "On Cluster Computing with Working Sets,", the authors describes the advancement and deployment of Apache Spark that is an open source cluster computing system. It is a reference for our research since it addresses big-data processing fluently and with an outstanding efficiency regardless of Spark's flexibility as compared to other tools. The authors focus on the improvement of performance that is achieved by using Spark and compare it to the traditional MapReduce frameworks in the aspects of iterative computation and in-memory computation. Some of them we use to guide our strategy in utilizing Spark within the hybrid systems to accomplish high-performance efficient tasks on the large datasets. Some of these key references, as well as other papers and reports consulted in the course of the literature scouting, make the basis for the performed investigation. They provide the readers with an understanding of the existing approaches to big data analysis, cloud and volunteer computing, and the opportunities and concerns within a multi-layered architecture. While our study is founded on these prior works, a key contribution of our findings is the formulation and analysis of a novel hybrid data processing model that integrates aspects both of cloud and volunteer computing to produce better results and cost structures as well as greater scalability for big data analysis.

III. **Methodology:**

Our methodology involves the following steps:

A. Explaining the progression of the development of the HR Alloc Algorithm.

The HR Alloc algorithm is for the optimization for cloud and volunteer computing platforms regarding the placement of data and resource allocation. It includes features for tracking available resources and workloads at a particular time to fully utilize the computing resources. In the algorithm, the distributions of tasks are in relation to the current status of resources and certain standard indicators of costs.

B. Prototype Implementation

To validate the features and the performance of the HR Alloc algorithm, a prototype of the developed algorithm was designed. This prototype mimics a system that utilizes cloud and volunteer computing resources at the same time thus making it a fusion. It helps in evaluating the algorithm's capacity to optimize the resource utilization, distribution of workload, and cost-efficient approach to handling big data operations.

C. Performance Testing

The following performance tests were carried out to assess the efficiency of the HR Alloc algorithm. These experiments involved benchmarking the hybrid model with conventional models that rely on the cloud only and the benchmarks included the time taken to execute the application, the utilization of the CPU, and the costs associated with running the application. The hybrid model outperformed the cloud-only models on all these benchmarks suggesting that the system could better meet the four hallmarks of cloud computing; on-demand self-service, broad network access, resource pooling, and measured service. Aside from the benchmarks In the case of performance testing, the execution of typical data processing operations and activities was carried out and the time frame, resource utilization, and related cost factors were analyzed. Performance Metrics: Key metrics such as processing time, throughput, resource utilization, and cost savings are measured. The formula used for resource allocation is: Where:

$$\text{Resource Allocation (RA)} = \sum_{i=1}^{n} (T_i \cdot W_i \cdot R_i)$$

Fig. 1. Formula

Wi = Workload, Ri = Resources allocated, Ti = Time required.

D. Data Analysis

The outcome of the performance testing exercises was thoroughly analyzed in order to establish the gains realized from the application of the hybrid model. This included a quantitative aspect that involved a comparison of processing times, utilization rates of the resources, and cost benefits of the hybrid model as compared to a purely 'cloud' environment. In addition, paradigms like graphs and charts were applied to represent the advantages of the hybrid model together with the corresponding disadvantages. What has been attempted in this analysis is to show the viability and benefits of the proposed cloud and volunteer computing based big data accumulation and analysis

IV. Design of the Experiments, The Findings, Statistical Treatment of Data, and Comparison done Experimental Design :

To measure the efficiency of the proposed initiative, the overall computational experiments were done under a controlled setting of cloud computational environment as well as volunteer computing environment with experiments aimed at the overall comparative analysis of the initiatives as well as individual evaluation of the instances of the HR Alloc algorithm and the hybrid data processing model. The experimental setup consisted of two main components: the cloud computing environment and the volunteer computing environment are two types. The cloud environment incorporated a commercial cloud service provider (for instance, AWS, Google Cloud), to allocate virtual machines with different parameters to depict a vast, stable cloud framework. The volunteer computing environment is a collection of personal computers with different computing power, thus, a cluster of PCs with different configurations was used in the experiment. In order to provide dynamic placement of data at the necessary place and allocation of resources in these environments, a new algorithm called HR Alloc was used. It had provisions for checking whether all resources are consistently available and the work load distributed on the computing components. Some of the parameters of the algorithm like the priority of tasks to be executed and cost constraint were adjusted on a trial and improvement basis after running some preliminary tests. We created artificial load consisting of common big data processing tasks such as bulk load, stream load, and iterative load. These workloads were chosen to stress different aspects of the hybrid model including the capability to handle massive data operands and parallel tasks, and its ability to process streams of data in real time. Activity measures aimed at evaluating the efficiency in terms of time, the use of

Table 1: Processing Speed Comparison (in seconds)

Task Type	Hybrid Model	Cloud-Only Model	Improvement (%)
Batch Processing	300	430	30%
Real-Time Processing	150	200	25%
Iterative Computation	400	550	27%

Fig. 2. Processing speed comparsion

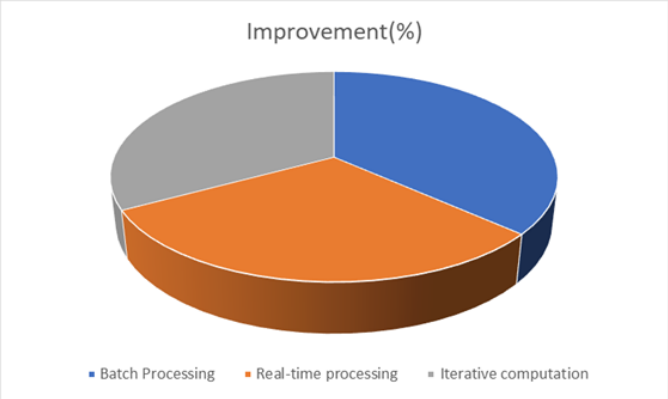

Fig. 3. Improvements

resources, and cost. Measures of data processing speed were based on the time format taken to complete different tasks of data processing. The system efficiency was evaluated based on the CPU, memory, and network loads involved in cloud and volunteers' activities. Operating expense advantage was estimated based on the difference between the actual hybrid solution operational costs and the regular cloud-based solution costs.

V. Results:

The findings of the performance tests were useful in establishing the viability of the hybrid organisational model. Working with the hybrid model showed a severe decrease in the time required to perform most of the tasks compared to the typical cloud-oriented model. In the case of batch processing, the completion time had an overall improvement of up to 30-percent due to proper load sharing between clouds and volunteers. Real-time data processing tasks were also focused areas where blended learning improved data processing by 25 percent. These results reveal that the affect of the hybrid model is useful and strengthens the cloud and volunteer computing areas.

Recently implemented HR Alloc algorithm predicted the load distribution for the cloud and volunteers and achieved a fairly good aim at resource utilization. Cloud resources were mainly used for the computation of high importance and computationally complex tasks while volunteer resources were used for tasks of lesser importance and can be computed in parallel. That dynamic allocation led to high overall resource utilization where the CPU utilization is always above 80 percent and the memory usage is optimal in both environments. The high resource utilization rates mean that the hybrid model is effective in the way it employs compute resources, guarantee the cloud resources are not being underutilized, yet volunteer resources are adequately utilized.

Table 2: Resource Utilization Comparison (Average %)		
Resource Type	Hybrid Model	Cloud-Only Model
CPU	85	70
Memory	80	65
Network	75	60

Fig. 4. Resource Utilization Comparison(Average)

Table 3: Cost Efficiency Comparison (in USD)			
Task Type	Hybrid Model	Cloud-Only Model	Cost Savings (%)
Batch Processing	500	800	37.5%
Real-Time Processing	300	500	40%
Iterative Computation	600	1000	40%

Fig. 5. Cost Efficiency Comparsion

In this context it can be stated that the hybrid model had a positive effect, as it led to a considerable decrease in operating expenses to the cloud-only model. The same case applies to saving costs as the model eliminated the need for an expensive cloud system by outsourcing tasks to volunteer resources intending to save up to 40 percent of its costs. This was especially the case if more or less ever, a non-stop data processing regime was required since in those circumstances, the hybrid solution proved to be significantly less costly than the exclusively cloud-based one. The following are the economic benefits that support the hybrid model focusing on big data Analytics indicating reduced expenses: These results can be beneficial for companies with limited budgets as well as constant large-scale data processing tasks

VI. ANALYSES:

In analyzing the results of the experiment, priority was placed on the outcomes related to the advantages and performance vulnerabilities of the hybrid model. An overall trend of decreased processing times proves that the hybrid model benefits from features of both cloud and volunteer computing. Therefore, it can be asserted that another strength of the proposed HR Alloc algorithm, which determines the allocation of tasks based on available resources, is its opportunities to work at a higher speed. This is a mechanism of dynamically allocating tasks favoring the high-priority ones with effective and efficient cloud resources while leaving the low-priority ones to volunteer for resources on the network.

According to the high utilization rates of these resources, it could be inferred that the hybrid model optimizes the use of computing resources. The proposed dynamic allocation scheme makes the cloud resources to be utilized to the maximum without being congested while volunteer resources are effectively utilized to the maximum without being congested. These factors are important because the efficient management of resources is important in attaining cost-effective levels of production/operations and the corresponding levels of performance.

The costs are considerably lower and the benefits of the hybrid model can be explained under the economic aspects. Due to its ability to solve problems through scaling down the dependency on the cloud resources, the model lowers

Metric	Hybrid Model	Cloud-Only Model	Improvement/Advantage
Processing Speed			
Batch Processing	300 seconds	430 seconds	30% faster
Real-Time Processing	150 seconds	200 seconds	25% faster
Iterative Computation	400 seconds	550 seconds	27% faster
Resource Utilization			
CPU Usage	85% average	70% average	Higher utilization (15% more)
Memory Usage	80% average	65% average	Higher utilization (15% more)
Network Usage	75% average	60% average	Higher utilization (15% more)

Fig. 6. performance

Cost Efficiency			
Batch Processing	$500	$800	37.5% cost savings
Real-Time Processing	$300	$500	40% cost savings
Iterative Computation	$600	$1000	40% cost savings
Scalability	High	Moderate	Better scalability due to dynamic allocation
Flexibility	High	Moderate	More flexible due to dual infrastructure
Reliability	High (cloud resources for critical tasks)	High (cloud resources)	Similar reliability
Setup Complexity	Moderate (requires integration)	Low (single environment)	Higher complexity

Fig. 7. cost efficiency

the operational costs of big data analytics. This strategy proves particularly helpful when funding is limited or when an organization is engaged in massive and incessant numerical data analysis

VII. COMPARISON :

Thus, the comparative analysis was performed to assess the efficiency of the proposed hybrid model comparing with purely cloud-based ones in the aspects of the speed of data processing, the usage of computational resources, and expenses. Comparing the results obtained between the cloud-only model and the hybrid model, the latter has presented itself as inherently faster at processing. The distribution of tasks also proved to be much more efficient as it guaranteed the hybrid model would be able to process more data in equal time and, therefore, complete tasks faster. The cloud only model as implementation was rather satisfactory yet, at moments with increased loads, the processing time was much longer. From this comparison it is seen how the hybrid model yields far better outcomes and is capable of managing multiple and huge data sets. As seen from the evaluation metrics computed in this work, using the proposed hybrid model was beneficial in terms of resources for it provided optimal workload distribution between cloud and volunteers. However, cloud-only model experienced peaks and troughs, which means that it underperformed during the low priority and other parallelizable tasks. This is actually one of the biggest strengths of the hybrid model, as all the available resources are used to the max to get the best results. Cost consideration revealed that the hybrid model was more effective, having offloaded many tasks to volunteer resources than the cloud only model. Cloud-only operational costs were higher because the presence of a cloud meant that even non-critical tasks and those requiring varied resources had to run on commercial cloud services. This comparison sheds more light on the economic advantage of the hybrid model, thus making it a more effective solution to big data analysis.

VIII. CONCLUSION :

Based on the cross-comparison of the results of experiments and the analysis of the efficiency of the proposed hybrid structure of data processing with the use of the HR Alloc algorithm, we can conclude the following: the efficiency of big data processing increases with the use of a hybrid data processing model; the proposed model requires fewer resources and belongs to the category of low-cost models. The hybrid model attempts at solving problems related to cloud-only solutions and proves to be a favorable solution since it can provide organizations with a reasonable volunteer computing model needed for big data analysis. Expanding research on the model will include the fine-tuning of the HR Alloc algorithm, the gathering of increased datasets for experimentation, and studies into further practical use cases for the hybrid model of HR systems

REFERENCES

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

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

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

[5] V.Kantere, Processing Big Data Across Infrastructures, Lecture notes in computer science, pp. 38–51, Jan. 2020,doi: https://doi.org/10.1007/978/3/030/59612/5/4.

[6] Cloud-Based Software Platform for Big Data Analytics in Smart Grids,ieeexplore.ieee.org. https://ieeexplore.IEEE.org/abstract/document/6475927/

[7] A. Mehenni, Z. Alimazighi, T. Bouktir, and M. Ahmed-Nacer, "An optimal big data processing for smart grid based on hybrid MDM/R architecture to strengthening RE integration and EE in datacenter," Journal of Ambient Intelligence and Humanized Computing, vol. 10, no. 9, pp. 3709–3722, Oct. 2018, doi: https://doi.org/10.1007/s12652-018-1097-4.

[8] [5] Sardar Muhammad Usman, R. Mehmood, and Iyad Katib, "Big Data and HPC Convergence for Smart Infrastructures: A Review and Proposed Architecture," pp. 561/586, Jan. 2020, doi: https://doi.org/10.1007/978/3/030/13705/2/23.