International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Balanced Prediction Based Dynamic Resource Allocation Model for Online Big Data Streams using Historical Data

Vijaya Kumar Chandarapu[2], Madhavi Kasa[2]

Submitted: 14/08/2022 **Accepted:** 23/11/2022

***Abstract*:** For cloud computing service providers, one of the largest problems is to maintain good service standards for resource allocation and distribution. The multitenant use of cloud resources with the help of cloud services is one of the most essential utilities in the current period of the technological world. End-user resources should be available with minimal administration and an efficient resource allocation strategy must be developed to avoid scenarios of overprovisioning or under provisioning of resources for handling big data streams. Clients want to keep their costs down, while cloud providers want to get the most out of their existing infrastructure while minimising the need for any new upgrades and provide service with minimum delay. Resource providers can take advantage of the elastic infrastructure provided by cloud computing to obtain streaming capabilities that precisely match demand of users. The amount of cloud resources allocated to users is billed to them. Cloud resource selection has been pioneered by the growing requirement to extract knowledge from massive data streams. The existing methods of resource allocation are dependent on the properties of the data themselves. Despite this, due of the random nature of data generation, it is impossible to predict the features of data in massive data streams. This presents a challenge in selecting and assigning resources to the stream of large data. This work presents a system that anticipates data qualities such as volume, velocity, variety, variability, and truthfulness in order to go in that direction. The proposed model considers weather forecasting data and classifies it as multiple tasks and resource allocation is performed for big data processing. The weather forecasting historical data is used by a set of server groups to assign distinct users jobs to the most trustworthy dynamic resources with less delay.This work presents an effective Balanced Prediction based Resource Allocation for Weather Streaming Data processing using Metadata (BPRA-WSD-MD). The proposed model takes weather forecasting streaming data as input and then divides the data into multiple jobs based on the historical weather report and performs the job execution done successfully by accurate resource allocation model. The proposed model is compared with the traditional models and the results represent that the proposed model performance levels in resource allocation is accurate.

Keywords: *Streaming Big Data, Historical Data, Dynamic Resource Allocation, Big Data, Minimum Delay, Clustering.*

1. Introduction

The term big data refers to a gathering of large amounts of semi structured and unstructured information to be processed by current data processing platforms [1]. Big data might be in the terabytes, petabytes, or zettabytes range in size [2]. The problem is that it is impossible to define clear thresholds for the volume of big data because it is a qualitative phrase. Due to increased storage capacity, it is possible that the volume of data that characterises big data [3] now may not satisfy the criteria in the future. Unstructured datasets may be termed huge data of a given size, whereas the same size structured dataset may not qualify [4]. As a result, the concept of big data has expanded beyond the volume of data to include other characteristics such as diversity and velocity [5]. Having a wide range of data kinds is known as diversity. Sensors, actuators, financial activities, social networks, other smart objects contribute to data diversity by capturing text, image, audio, even video data [6]. There is a large amount of data in an unstructured form. However, due of its rapid pace of generation, even a modest amount of data can be termed big data. There are no fixed data velocity thresholds because of these issues.

There are additional characteristics of big data, such as validity, variability, value, and visualisation in addition to volume, variety, and velocity. In IBM's coining of the term "veracity," they were referring to the unreliability of some data sources. Most big data comes from social media, corporate transactions, stock markets, weather forecasting centres and cyber-physical systems [7]. Big data is constantly being generated by these apps all around the world. When data travels so quickly that it cannot be stored in its whole, it is considered as streaming data that updates regularly [8]. When the application requires immediate response to data, it is also considered streaming data [9]. For instance, data is generated at a very high rate via social media or by weather forecasting centres. All of the above-

[1]*Research Scholar, Department of Computer Science and Engineering Jawaharlal Nehru Technological University College of Engineering, Jawaharlal Nehru Technological University, Ananthapuramu - 515002, Andhra Pradesh, India.*
vijay.chandarapu16PH0524@gmail.com
[2]*Associate Professor, Department of Computer Science and Engineering Jawaharlal Nehru Technological University College of Engineering, Jawaharlal Nehru Technological University, Ananthapuramu - 515002, Andhra Pradesh, India.*
kasamadhavi@yahoo.com

mentioned key sources of data can be thought of as large data in this context.

Knowledge extraction from massive amounts of data is both necessary and difficult. A cloud-based solution is required since it goes beyond the limits of traditional on premise solutions [10]. The selection of the best number of cloud capabilities for each user requests has arisen as a critical difficulty the with growth of data on cloud [11]. In addition, the need for real-time analysis of large data streams makes it even more critical to choose the right resources [12]. In order to achieve this goal, it is proposed that two methods for identifying acceptable resources should be used. Users can pick and choose which resources to employ in the first method. The user must be familiar with the data's qualities in order to make such an allocation [13]. When dealing with large amounts of data, the user may opt for a larger memory capacity, a more powerful GPU, or a greater processing power to accommodate the increased speed and variability [14]. As a result, the user's level of knowledge dictates the resources availability. Furthermore, even if the user has sufficient expertise, it is vital for them to be familiar with data characteristics [15]. However, the properties of an incoming big data streams are often unpredictable due to the dynamic nature of data production from multiple sources [16]. Because of this, selecting cloud resources by hand isn't the best option for handling large amounts of data that will increase the load on the cloud system because of lack of resources or with inefficient resource scheduling models. The cloud resource pool and VM that are allocated with the tasks are shown in Figure 1.

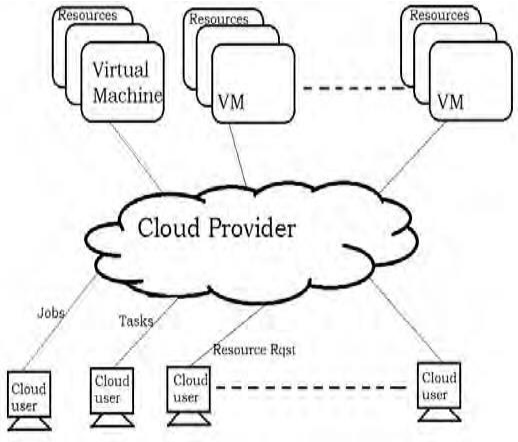

Figure 1: Resource Pool and VM with Tasks

At some points in time, forecasting models can anticipate what the weather will be like using a computer programme [17]. Modern forecasting models use Numerical Weather Prediction Analysis (NWPA), which is defined as a system of reduced equations used to regularly revised in atmospheric conditions. Cloud computing gives scientists new alternatives for data processing and modelling simulations by utilising cloud-deployed software, information sharing and collaboration system, and cloud-based computer resources [18]. Using IaaS concepts to improve regional numerical weather prediction, this research examines cloud terminology relevant to the meteorological community [19]. Because developing countries may not have access to conventional supercomputing facilities, they are given special attention. Depending on the layout of Amazon Elastic Compute Cloud (EC2) resources, regional weather simulations cost $40 to $75 each 48-hour forecast. Cloud service providers can benefit from reliable forecasting of future demand for a cloud resource by using historical data trends. With dynamic scaling of cloud resources, predicting resource demand is essential to reaping the benefits, such as cost reductions and optimal energy consumption [20].

Data for training and testing are essential in any machine learning task. The data created by any online application is always random, but when associated to calendar features, they tend to indicate a general trend and seasonality in most circumstances. We may need to add current data to historical data if fundamental changes in the system being predicted that make previous data less valuable [21]. In this scenario, we can improve our model while keeping usable data for a longer period. Attempting to predict the future based on the past is an example of extrapolation in a weather forecasting problem. Models are built for the given data, and then used to fit further data that is not within that range. The model needs to be fine-tuned if we receive a new forecast that alters previously-known data, which is another consideration when making a weather forecast [22]. It should be also considered that predicting over a longer period of time is prone to error. This research propose a Balanced Prediction based Resource Allocation for Weather Streaming Data processing using Metadata for efficient resource scheduling.

2.Literature Survey

The term big data was coined by Rayet al.[1] and has since been traced back to a variety of origins. A big data era has begun as a result of the explosion of data generated by applications like social networks, e-commerce transactions, mobile app, cyber-physical system and IoT. Big data is gaining some traction as a result of this spike in popularity. There have been numerous summaries of the current state of big data analysis by a variety of authors. With the use of cloud computing, big data analytics may be made more efficient. That cloud provides essential support for big data and analytics by leveraging multiple processing and storage resources supports. Nayak et al.[3] claimed that cloud-based big data are on the rise. They said that one of the most difficult aspects of working with large amounts of data is figuring out which methodologies and cloud resources to use. Tarafdar et al. [5] proposed that the appropriate number of cloud services should be decided prior to the commencement of a project's actual implementation. It was as a result that the field of cloud resource prediction and provisioning gained prominence. In spite of this, research on scheduling large data applications is still lacking. For big data applications, dynamic resource provisioning is a difficult problem to solve. Peng et al. [6] suggested a number of research directions for real-time, global, dynamic, adaptive, and multi-objective programming of big data applications.

Fatima et al. [8] stated that for smoother scaling performance, users need to employ the proactive technique of anticipating resource metrics and using it to provision future cloud resources in order for load balancing. Support Vector Machine (SVM), Neural Networks (NN), and Linear Regression (LR) were all discussed by Khodak et al. [10]. The SVM was found to be the most effective method for ML, and it was recommended. New approaches to resource burden prediction based on previous consumption patterns were discussed in the model proposed by Yin et al. [11] Dynamic resource allocation has a lot of overhead in terms of responsiveness, setup time,

etc. Scaling overhead can be reduced if the system can forecast and adapt to the incoming execution task needs, according to the work done by Agarwal et al.[12]. They developed a pattern matching algorithm based on similar characteristics of online traffic to identify patterns most closely related to resource utilisation metrics in the cloud's historical data. This method, however, has few drawbacks because it incurs a computational cost every time a search for similar patterns in the historical data is conducted. Overfitting, on the other hand, is something we want to avoid at all costs.

Lu et al. [14] suggested a method for forecasting virtual machine resource requirements based on projected application workloads. To anticipate the resources needed for applications, Patel et al. [17] employed Artificial Neural Networks (ANN) and Linear Regression (LR). In contrast to ARIMA, which relies on the complicated representation of time series and ANN-based algorithms to effectively predict, simpler techniques such as linear regression can yield predictions faster, but they also need that metrics have more straightforward behaviour. According to Chen et al. [18], a time series-based model can be used to efficiently to retrieve present action from centre workloads. Because of the non-stationarity element in time series data, users agree to use ARIMA prediction for the architecture and compare it with machine learning-based neural network model in order to increase predictive model's applicability.

Gurleen et al. [23] emphasised that the resource scheduler has a significant impact on big data applications. According to their research, the best routing protocol can be identified under certain circumstances. When it comes to scheduling big data solutions over dispersed clouds, Laili et al. [24] suggested a quality of the service architecture. Big data began to flow in continuous streams as the Internet became more widespread and widely used. The issue of extracting meaningful insights from massive amounts of data is widely acknowledged. Big data streams can be mined using a variety of algorithms and strategies described by various authors. Big data streams can be collected, integrated, analysed, and visualised in real time using the provided algorithms. Despite this, resource scheduling for massive data streams has received less attention than it should.

It was proposed by Kavitha et al. [26] that a graph-based strategy be used to efficiently manage the resources stored by stream processing tools. A strategy for processing data streams in such a way that the cloud provider's profit is maximised. Using a priority-based approach, Vahidi et al. [27] devised a strategy for handling multimedia data. Motlagh et al.[31] used Markov chains to accurately anticipate the volume of big data and then assigned node for data processing in accordance with their findings. In addition to the Vs of big data and resource allocation, the approach does not take into account another Vs of massive data.

3. Proposed Model

Many services can be supported by cloud computing's vast compute, storage and network capabilities [23]. Besides being able to access resources and services at any time and from any location, users can also increase or decrease the number of resources they use to optimise their applications' performance or reduce their overall costs. Cloud resource requests are becoming more diverse, sporadic, and sudden as a result of the rise of artificial intelligence and machine learning [24]. For a high number of abrupt resource requests, the current cloud allocation of resources methods can't ensure timely and optimal resource allocation. While cloud service providers are more concerned with how to manage the huge resources and enhance resource utilisation, users give greater attention to the timing and optimization of their emergence resources and demand that can ensure their applications' performance. These objectives necessitate an effective means of allocating resources [25].

Based on quantity, velocity, variety, validity, and variability, the suggested system tries to distribute suitable resources to huge data streams. Data from a large number of weather forecasting historical data sources are processed on the cloud to produce desired results for accurate future weather predictions. Here, it is observed that the big data stream of historical weather data is made up of a wide range of data items. The suggested system selects relevant resources for studying a cloud stream made up of multiple weather task items. The suggested system utilises two modules, namely the Task Scheduling and the Resource Management to predict weather forecasting based on historical data. The proposed model framework is shown in Figure.

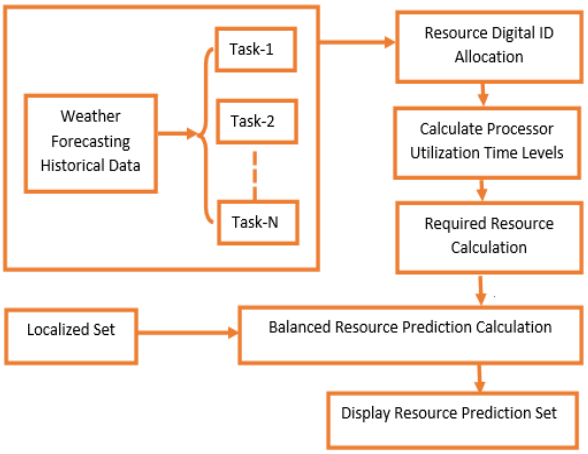

Figure 2: Proposed Model Framework

In large-scale distributed systems, such as cloud computing, resource provisioning is critical. More than one system for allocating resources to ensure high levels of service quality used pre-emptive resource allocation. A lack of resources for high-priority requests can be alleviated by delaying or cancelling lower-priority leases or jobs to free up the resources needed for the higher priority requests. This paper briefly discusses how to effectively allocate processing resources for weather streaming data on the basis of balanced predictions. Using taking weather forecasting streaming information and dividing it into several jobs based on past weather reports, our suggested approach successfully conducts task execution by an accurate resource allocation model.

Algorithm BPRA-WSD-MD

{

Input: Weather Data as Tasks for numerous server groups.

Output: Weather Prediction Set based on Historical data experiences.

Step-1: Create a set of tasks that are divided from the historical weather forecasting data and then assign the tasks to the virtual machines. The tasks are listed in the job queue as

Task-List[N] = {T_1, T_2, …… , T_M}

Step-2: The resources which are provided by the cloud service provider are allocated with the digital identities (Resource Digital Identity-RDI) for handling and revoking them accurately for effective resource scheduling. The resource identity allocation is done and maintained as a set in the resource pool as

RDI = {Res_1:RDI_1 , Res_2:RDI_2, …… , Res_N:RDI_N}

Step-3: The tasks that are divided from the weather historical data are calculated with the total processor utilization time based on the length of instructions. The processor utilization time levels of each task is calculated as

$$\text{Pr}ocess-Util[T(N)]=\frac{\sum_{i=1}^{N}Task\_len(_{T)}+sizeof(T)}{\sum_{i=1}^{N}size(VM)}+\frac{\sum_{i=1}^{N}\max len(T(i)+Th+\max tasks(VM(i))}{sizeof(T)}$$

Step-4: The required number of resources for execution of a task is calculated and the resources are scheduled from the resource pool. The required number of resources for each task varies depend on the processor utilization time and length. The resource requirement is calculated is performed as

$$\text{Re}s\_\text{Re}q(T[N])=\frac{\sum_{i=1}^{N}Task\_len(T)+\max(\text{Pr}ocess-Util(T(i))}{\sum_{i=1}^{N}\sqrt{\frac{VM(i)+\max(T)_i^{Length})}{sizeof\ (T(i))}}}+\sum_{i=1}^{N}\sqrt{\frac{MaxLimit(VM))}{Alloc-Task(VM(i))}}$$

Step-5: Based on the resource requirements of tasks, the resources are allocated to the tasks that are allocated to a virtual machine group. The resources are allocated from the resource pool that can be tracked using the RDI. The resource allocation is performed as

$$Ralloc(T(i)\varepsilon VM(i))\ =\sum_{i=1}^{M}\max(\text{Re}s\_\text{Re}q(T(i))+RDI(T(i))+\left(\frac{(MaxLimit(VM))}{\|ToTal\_Tasks(N)\varepsilon VM(i)\|}+\max tasks(VM(i))\right)$$

Step-6: The historical data is analysed based on the tasks that are allocated to the VMs. The data is considered and weather forecasting models are stored in the localized set for further utilizations. The localized set is generated as

$$Localized-Set(T(i))$$
$$=\sum_{Task=0}^{N}\frac{Task\_len(T(i))+\sum_i^M RDI(T(i))}{\text{MaxLimit(VM)}}$$
$$+\sum_{i,j}\frac{|\text{getattr(T(i)}|}{\text{sizeof(T(i)}}$$

$$Balanced\ \text{Re}s-alloc[T(i))=\frac{\sum_{i=1}^{M}getRDI(VM(i))+Task\_len(T(i))+\max Tasks(VM(i)}{\Sigma_{i=1}Totaltasks(VM(i))}\qquad\frac{balanced\ if\ \ getRDI<Th}{otherwise\ \ Ralloc}$$

Step-7: The resource balanced prediction is applied on the localized set for scheduling the resources to all the weather historical tasks so that the localized set will be updated for accurate weather predictions. The resource balanced prediction allows to reduce the load on the network by avoiding delay in the VM group. The resource balanced prediction is done by allocating the resources uniformly to the tasks for weather predictions. The balanced prediction is performed as

Step-8: The weather data load prediction on the VMs are calculated and then the resource utilization levels are observed based on the load. The load prediction on a VM is calculated as

$$Load(VM(N))=\max Tasks(VM(i))+\min(Ralloc(T(i))-Localized-Set(T(i))+\frac{\max len(T(i))}{sizeof\ (VM(i))}$$

Step-9: The further weather forecasting predictions can be done using the localized set and the prediction set is generated based on the historical weather data analysis as.

$$Pred-Set(VM[N])=$$
$$\frac{\sum_{i=1}^{\text{M}}\text{getRDI(T(i))+getattr(Localized}-\text{Set(T(i))+max len(VM(T(i))}}{\sum_{i=1}^{\text{N}}Total-Task(T(i))}+$$
$$\min\ (load(VM(T(i))))$$
}

4. Results

As a consequence of climate change and other climatic abnormalities, extreme weather conditions are becoming more common in many nations, resulting in significant damage and death. Conventional weather forecasting systems can't provide predictions for such wide areas over such a long timescale, thus they're useless in this situation. Localized sensors linked to cloud computing can deliver real-time forecasts for tiny locations in numerous weather prediction systems. There are many types of time series data, but the most common is referred to as a Weather based Time Series Data. If we apply this definition to our case, we may say that a weather time series data is one in which the observations are made at regular intervals of time (TI = t1, t2, t3,..., TN). We have high-frequency data at regular intervals in time, whereas most time axes are uniformly spaced across a period of hours, days, or even years, which are referred to as low-frequency data.

The proposed effective Balanced Prediction based Resource Allocation for Weather Streaming Data processing using Metadata (BPRA-WSD-MD) is implemented in cloud simulator and the proposed model considers the weather forecasting historical data from the link https://www.kaggle.com/mahendran1/weather-data-in-india.

The proposed mdoel is compared to the traditional Resource Allocation Based on Predictive Load Balancing Approach in Multi Cloud Environment (RA-PLB-MCE) Model. The proposed model is compared with the traditional models in terms of Historical data analysis time levels, Resource ID allocation Accuracy levels, Task Classification Time Levels, Resource Allocation Time Levels, Resource Failure Handling Accuracy Levels, VM migration and resource handling accuracy levels, Resource Utilization Prediction accuracy levels, Task-Resource Scheduling Accuracy Levels, Weather Forecasting Accuracy Levels.

The study of weather behaviour over a specific period of time is called historical weather data analysis. When discussing the many diverse aspects of the weather conditions, the predictions can be done. A statistical edge can be identified and built for active prediction by analysing previous data. The historical weather data analysis time levels of the proposed and traditional models are shown in Figure 3. The proposed model in less time can analyse the historical data that can be used for further predictions.

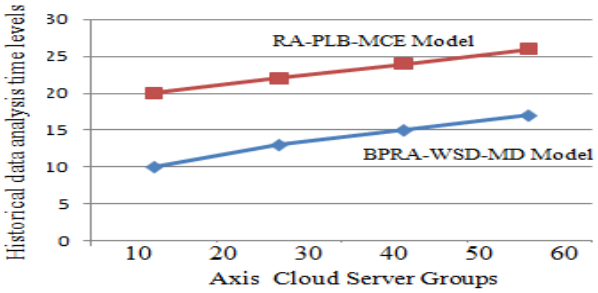

Figure 3: Historical Data Analysis Time Levels

The proposed model initially registers all the resources that are in the resource pool for effective scheduling to the weather tasks. Based on the resource registrations identities, the resources are further allocated to the tasks for execution. The resource identity allocation of the proposed and traditional models is shown in Figure 4.

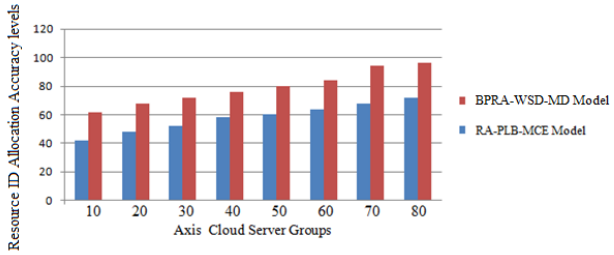

Figure 4: Resource ID Allocation Accuracy levels

In cloud computing, scheduler is the key issue that impacts the system performance. An effective task-scheduling method is required to increase system performance. Existing task-scheduling methods concentrate on the needs of task-resources, CPU memory, implementation time and cost. Task scheduling refers to the process of selecting which task will take up the most processor time. The order in which the tasks are completed may be determined by their importance and time of execution. The proposed model allocates the resources based on task execution time. The task classification time levels of the proposed and traditional models are shown in Figure 5.

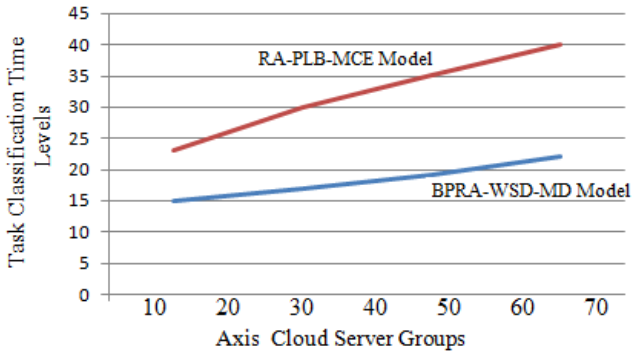

Figure 5: Task Classification Time Levels

Asserting and managing resources in accordance with an organization's strategic goals is known as resource allocation. The term resource scheduling refers to the process of allocating resources among cloud users in accordance with a set of rules and regulations. Cloud computing relies on a fundamental technology called resource scheduling, which is used to manage resources. The resource allocation time levels of the proposed and traditional models are shown in Figure 6.

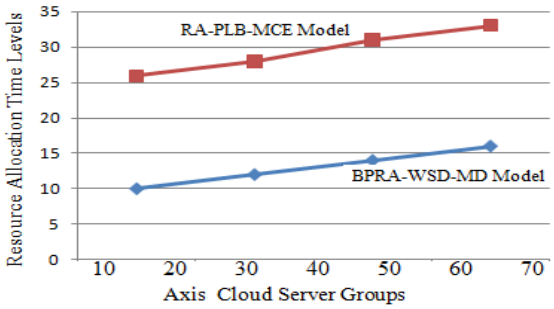

Figure 6: Resource Allocation Time Levels

When the resource detracts from its typical course of operation without meeting the requirements of the task, resource failure results. The resource failure will the tasks enter into waiting state and delay the execution. The proposed model effectively handles the failed resources and allocates new resources to the tasks. The resource failure handling accuracy levels of the proposed and traditional models are shown in Figure 7.

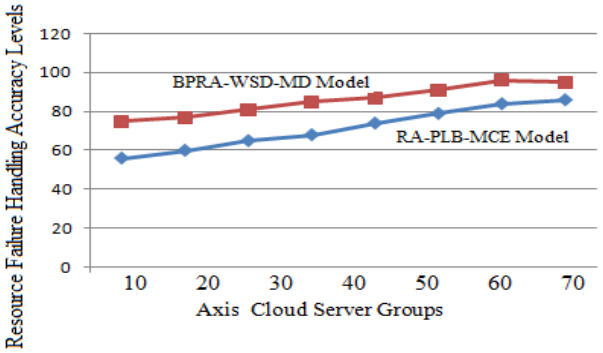

Figure 7: Resource Failure Handling Accuracy Levels

The process of migrating a virtual machine from one real hardware configuration to another is known as virtual machine migration. It's an aspect of hardware virtualization management and a consideration for service providers who offer virtualization. The VM migration and resource handling

accuracy levels of the proposed and traditional models are shown in Figure 8.

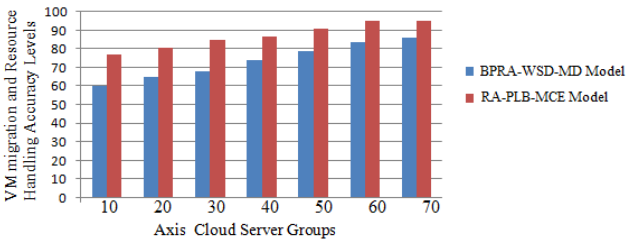

Figure 8: VM migration and Resource Handling Accuracy Levels

Optimal resource provisioning necessitates the use of a resource use prediction method. In a dynamic resource usage environment, it's difficult to make precise predictions. The proposed model predicts the resource utilization levels efficiently for task execution. The resource utilization prediction accuracy levels of the proposed and traditional models are shown in Figure 9.

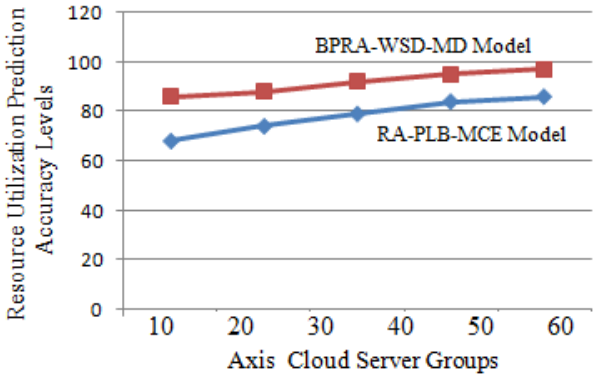

Figure 9: Resource Utilization Prediction Accuracy Levels

The resources are balanced among the tasks in the queue to avoid delays in the cloud environment. The resources are allocated to the tasks with balanced levels such that not tasks should wait for lack of resources and the execution should be effective. The Task-Resource balancing accuracy levels of the proposed and traditional models are shown in Figure 10.

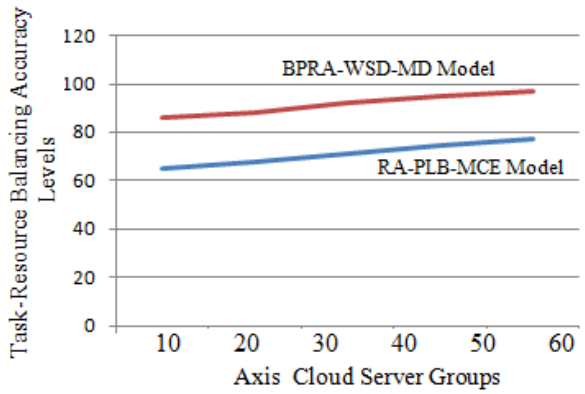

Figure 10: Task-Resource Balancing Accuracy Levels

Assessment and analysis, extrapolate to determine the future condition of the environment, and prediction of specific factors all go into making a weather forecast. The proposed cloud environment considers historical weather forecasting data and considers them as tasks and perform resource allocation for execution for accurate weather predictions. The weather forecasting accuracy levels of the proposed and traditional models are shown in Figure 11.

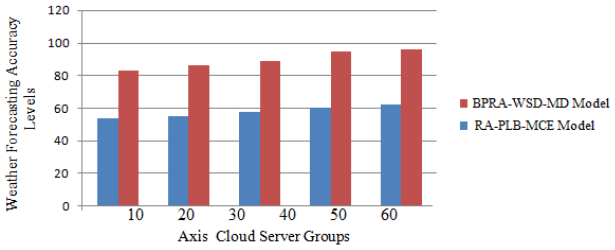

Figure 11: Weather Forecasting Accuracy Levels

5. Conclusion

With a real-time dynamic solution, it is possible to predict the volume, velocity, variety, variability, and truthfulness of big data streams in real time. The use of these 5Vs makes it possible to dynamically distribute cloud resources to large data streams. Software, data sharing and collaboration systems, and the utilisation of cloud computing infrastructure for data processing models are all new options for the scientific community in cloud computing. Using infrastructure as a service (IaaS) concepts as a foundation for regional numerical weather prediction, this research proposed a dynamic resource allocation model for weather predictions where weather historical data is provided as tasks and these tasks are allocated with resources for completing their execution for predictions. The load in the server groups need to be monitored frequently to avoid delays and failures of VMs. High-performance computing principles in the meteorological industry and in particular numerical weather prediction are easily aligned with cloud-based support for IaaS. This research presents an effective Balanced Prediction based Resource Allocation for Weather Streaming Data processing using Metadata. The resource scheduling is done effectively avoiding load on the cloud environment. In future, the power consumption can be considered as a factor that can be reduced in the cloud environment during resource utilization to improve the system efficiency.

References

[1]. Ray, B., Saha, A., Khatua, S, Roy, S.: Proactive fault-tolerance technique to enhance reliability of cloud service in cloud federation environment. IEEE Transactions on Cloud Computing (2020)

[2]. Maurya, A.K., Modi, K., Kumar, V., Naik, N.S., Tripathi, A.K.: Energy-aware scheduling using slack reclamation for cluster systems. Clust. Comput. 23(2), 911–923 (2020)

[3]. Nayak, S.C., Tripathy, C.: Deadline sensitive lease scheduling in cloud computing environment using ahp. J. King Saud Univ. Comput. Inf. Sci. 30(2), 152–163 (2018)

[4]. Ray, B.K., Saha, A., Khatua, S., Roy, S.: Toward maximization of profit and quality of cloud federation: solution to cloud federation formation problem. J. Supercomput. 75(2), 885–929 (2019)

[5]. Tarafdar, A., Debnath, M., Khatua, S., Das, R.K.: Energy and quality of service-aware virtual machine consolidation in a cloud data center. J. Supercomput., 1–32 (2020)

[6]. Peng, Z., Lin, J., Cui, D., Li, Q., He, J.: A multi-objective trade-off framework for cloud resource scheduling based on the deep q-network algorithm. Clust. Comput. 23, 2753–2767 (2020)

[7]. Ashraf, A., Porres, I.: Multi-objective dynamic virtual machine consolidation in the cloud using ant colony system. Int. J. Parallel Emergent Distrib. Syst. 33(1), 103–120 (2018)

[8]. Fatima, A., Javaid, N., Anjum Butt, A., Sultana, T., Hussain, W., Bilal, M., Akbar, M., Ilahi, M., et al.: An enhanced multi-objective gray wolf optimization for virtual machine placement in cloud data centers. Electronics 8(2), 218 (2019)

[9]. Jia, R., Yang, Y., Grundy, J., Keung, J., Li, H.: A deadline constrained preemptive scheduler using queuing systems for multi-tenancy clouds. In: 2019 IEEE 12th International Conference on Cloud Computing (CLOUD), pp. 63–67. IEEE (2019)

[10]. Khodak, M., Zheng, L., Lan, A.S., Joe-Wong, C., Chiang, M.: Learning cloud dynamics to optimize spot instance bidding strategies. In: IEEE INFOCOM 2018-IEEE Conference on Computer Communications, pp. 2762–2770. IEEE (2018)

[11]. Yin, L., Luo, J., Luo, H.: Tasks scheduling and resource allocation in fog computing based on containers for smart manufacturing. IEEE Trans. Industr. Inf. 14(10), 4712–4721 (2018)

[12]. S. Agarwal, F. Malandrino, C. F. Chiasserini and S. De, "VNF Placement and Resource Allocation for the Support of Vertical Services in 5G Networks", IEEE/ACM Transactions on Networking, vol. 27, no. 1, pp. 433-446, Feb. 2019.

[13]. L. Yala, P. A. Frangoudis, G. Lucarelli and A. Ksentini, "Cost and Availability Aware Resource Allocation and Virtual Function Placement for CDNaaS Provision", IEEE Tran. on Network and Service Management, vol. 15, no. 4, pp. 1334-1348, Dec. 2018.

[14]. Lu, L., Yu, J., Zhu, Y., Li, M.: A double auction mechanism to bridge users' task requirements and providers' resources in two-sided cloud markets. IEEE Trans. Parallel Distrib. Syst.. 29(4), 720–733 (2018)

[15]. Zhang, J., Yang, X., Xie, N., Zhang, X., Vasilakos, A.V., Li, W.: An online auction mechanism for time-varying multidimensional resource allocation in clouds. Future Gener. Comput. Syst. 111, 27–38 (2020)

[16]. Middya, A.I., Ray, B., Roy, S.: Auction based resource allocation mechanism in federated cloud environment: TARA. IEEE Transactions on Services Computing (2019)

[17]. Patel, Y.S., Nighojkar, A., Misra, R.: Truthful double auction based vm allocation for revenue-energy trade-off in cloud data centers. In: Proceedings of the 2019 National Conference on Communications (NCC), Bangalore, India, pp. 1–6 (2019)

[18]. Chen, J.X. Lin, Y. Ma et al., Self-adaptive resource allocation for cloud-based software services based on progressive QoS prediction model. Sci. China Inf. Sci. 62(11), 1–3 (2019)

[19]. J. Chen, Y. Wang, A resource request prediction method based on EEMD in cloud computing. Proc. Comput. Sci. 131, 116–123 (2018)

[20]. J. Chen, Y. Wang, A hybrid method for short-term host utilization prediction in cloud computing. J. Electr. Comput. Eng. 2782349, 1–14 (2019)

[21]. D. Shen, Research on application-aware resource management for heterogeneous big data workloads in cloud environment. Dongnan University, 2018.

[22]. X. Chen, J. X. Lin, B. Lin, T. Xiang, Y. Zhang and G. Huang, Self-learning and self-adaptive resource allocation for cloud-based software services. Concurrency Comput. Pract. Exp., 31(23), e4463 (2019).

[23]. K. Gurleen, B. Anju, A survey of prediction-based resource scheduling techniques for physics-based scientific applications, Mod. Phys. Lett. B, 32(25), 1850295(2018).

[24]. Y.J. Laili, S.S. Lin, D.Y. Tang, Multi-phase integrated scheduling of hybrid tasks in cloud manufacturing environment. Robot. Comput. Integr. Manuf. 61, 101850 (2020)

[25]. K. Reihaneh, S.E. Faramarz, N. Naser, M. Mehran, ATSDS: adaptive two-stage deadline-constrained workflow scheduling considering run-time circumstances in cloud computing environments. J. Supercomput. 73(6), 2430–2455 (2017)

[26]. K. Kavitha, S. C. Sharma, Performance analysis of ACO-based improved virtual machine allocation in cloud for IoT-enabled healthcare. Concurr. Comput. Pract. Exp., e5613 (2019).

[27]. J. Vahidi, M. Rahmati, in IEEE 5th Conference on Knowledge Based Engineering and Innovation (KBEI). Optimization of resource allocation in cloud computing by grasshopper optimization algorithm, pp. 839–844 (2019).

[28]. U. Rugwiro, C.H. Gu, W.C. Ding, Task scheduling and resource allocation based on ant-colony optimization and deep reinforcement learning. J. Internet Technol. 20(5), 1463–1475 (2019)

[29]. S. Shenoy, D. Gorinevsky, N. Laptev, Probabilistic Modelling of Computing Request for Service Level Agreement. IEEE Trans. Serv. Comput. 12(6), 987–993 (2019)

[30]. Z.H. Liu, Z.J. Wang, C. Yang, Multi-objective resource optimization scheduling based on iterative double auction in cloud manufacturing. Adv. Manuf. 7(4), 374–388 (2019)

[31]. A. Motlagh, A. Movaghar, A. M. Rahmani, Task scheduling mechanism in cloud computing: a systematic review. Int. J. Commun. Syst. e4302 (2019).

[32]. M. Kumar, S.C. Sharma, A. Goel, S.P. Singh, A comprehensive survey for scheduling techniques in cloud computing. J. Netw. Comput. Appl. 143, 1–33 (2019)

[33]. N. D. Vahed, M. Ghobaei-Arani, A. Souri, Multiobjective virtual machine placement mechanisms using nature-inspired metaheuristic algorithms in cloud environments: a comprehensive review. Int. J. Commun. Syst. 32(14), e4068 (2019).

[34]. F. Sheikholeslami, N. J. Navimipour, Auction-based resource allocation mechanisms in the cloud environments: a review of the literature and reflection on future challenges. Concurr. Computat. Pract. Exp., 30(16), e4456 (2018).

[35]. G. Natesan, A. Chokkalingam, An improved grey wolf optimization algorithm based task scheduling in cloud computing environment. Int. Arab J. Inf. Technol. 17(1), 73–81 (2020)

[36]. M. A. Reddy, K. Ravindranath, Virtual machine placement using JAYA optimization algorithm. Appl. Artif. Intell. https://doi.org/10.1080/08839514.2019.1689714.

[37]. S. Souravlas, S. Katsavounis, Scheduling fair resource allocation policies for cloud computing through flow control. Electronics 8(11), 1348 (2019).

[38]. L. Guo, P. Du, A. Razaque, et al. IEEE 2018 Fifth international conference on software defined systems (SDS). Energy saving and maximize utilization cloud resources allocation via online multi-dimensional vector bin packing (2018), pp. 160–165.

[39]. N. Gul, I. A. Khan, S. Mustafa, o. Khalid, A. U. R. Khan, CPU-RAM-based energy-efficient resource allocation in clouds. J. Supercomput. 75(11), 7606–7624 (2019).

[40]. R.L. Sri, N. Balaji, An empirical model of adaptive cloud resource provisioning with speculation. Soft. Comput. 23(21), 10983–10999 (2019)