Cost-Aware Infrastructure Automation Using Predictive Analytics for Multi-Cloud Environments

Adithya Jakkaraju

1. Abstract

The focus of this research paper is to use cost aware infrastructure automation on predictive analytics in Cloud computing environments. It studies the application of predictive analytics for allocating the scarce resources, minimizing the cost and improving the performance. It studies several cloud management tools with AWS as a case study. AWS's services Auto Scaling, Lambda, and Amazon Forecast help businesses to scale efficiently, manage costs easily and automate operations. Predictive analytics changes the game of cloud infrastructure managing: it is able to automatically adjust the resource dynamically and predictively managing costs. In the competitive digital space, this makes sure businesses need scalability, flexibility, and cost efficiency.

2. Keywords

Cloud, AWS, Automation, Cost

3. Introduction

The rapid growth of cloud computing adds to the business' new challenges of "cost manage and resource efficiency". With increasing complexity of cloud environments, traditional approaches of resource allocation and cost management is not efficient.

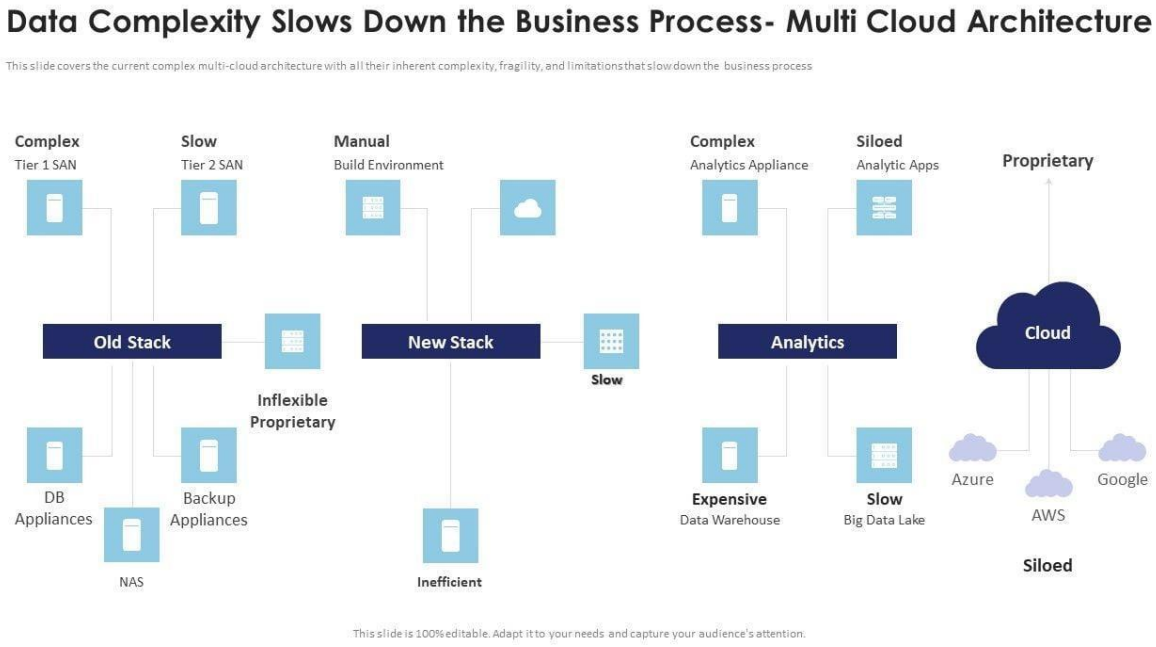

Figure 1 Multi cloud architecture (Slidegeeks, 2023)

This issue leads us to predict a solution in the form of predictive analytics, which combined with data driven insights will help us intelligently make decisions. Businesses can optimize the performance while minimizing the cost by forecasting their workload needs and then automating the resource scaling.

In this research paper, the impact of predictive analytics is studied on cost aware infrastructure automation bearing in mind amazon web services (AWS) as a practical example. Those tools are assessed in the study and determined to have potential for predicting cost cost and to perform in the real world.

4. Related works

4.1 Cost Optimization

Lately, cloud computing is serving as the integral part of business operations as the solution provides scalability, flexibility, and cost efficiency. But with organizations taking on more and more of a multi cloud environment, the ability to manage costs has become a more and more pressing problem.

Chinamanagonda (2020) claims that with the advent of cloud technologies, the complexity of cost management has heightened, though organizations are trying to optimally achieve a good performance in tandem with financial sustainability. Cost optimization does not mean just cutting expenses, but replacing resources with business needs by strategic use of monitoring tools and automation to achieve efficiency.

However, AWS, as a primary player in the cloud computing field, provides a great number of services and pricing models, which can result in cost management issues unless carefully handled (Tatineni, 2019).

Optimizing cost in AWS includes choosing the appropriate instance size, making use of reserved and spot instances as well as managing services in the AWS managed services. Additionally, managing costs in a multi cloud environment depends upon the adoption of a hybrid approach – that can be had with AWS and other cloud markets.

Businesses can allocate resources dynamically, to keep costs as they should be, instead of overprovisioning, by integrated predictive analytics and workload forecasting.

4.2 Predictive Analytics

The cost optimization in the cloud has therefore led to the revolution of predictive analytics for predicting the workload and automating resource allocation to manage workloads properly. As pointed out by Pothu and Kailasam (2023), workload prediction is one of the most essential factors for utilizing hybrid cloud environments in an efficient manner.

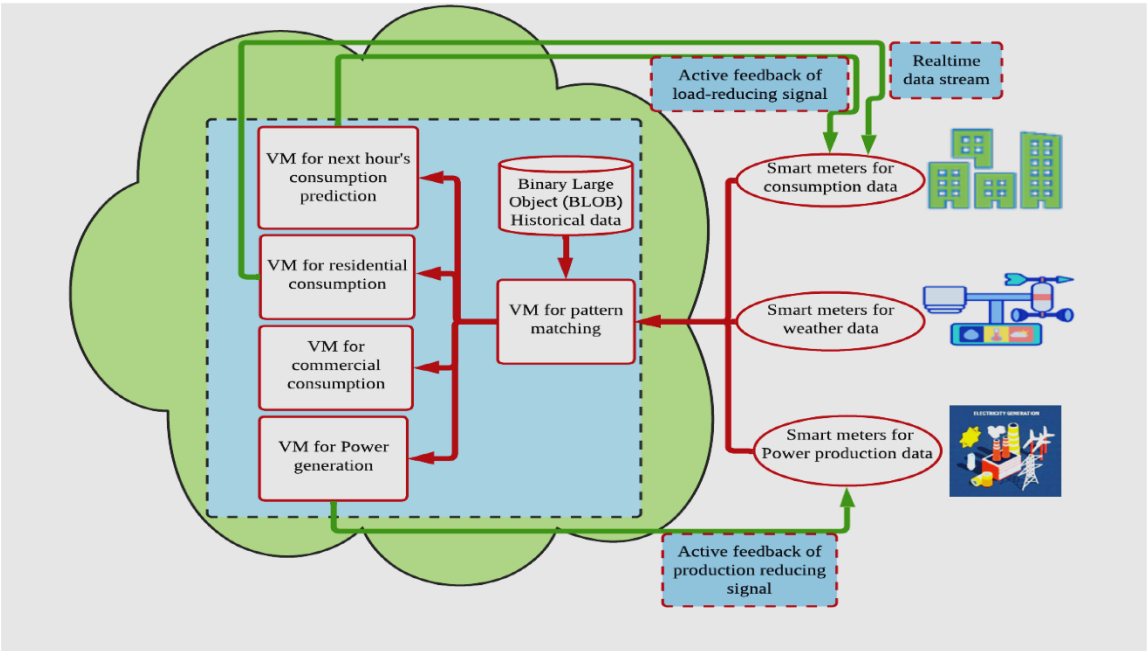

Figure 2 Multi cloud power management system (MDPI, 2023)

Often workload demands are predicted by machine learning and time series models, but they Common failure mode is when a radical change occurs. Enhancing predictive accuracy in Markov models and reinforcement learning also has been demonstrated but generally require enormous computational resource.

These models also allow for dynamic scaling and so that resources are being efficiently used without any cost. As also mentioned by Aldossary (2021), for instance, the integration of predictive analytics with AWS cost management should not be restricted to workload prediction for inclusion of energy aware optimizations.

Such a comprehensive approach to energy consumption and related costs reduction can be achieved by implementing predictive mechanisms at the Physical Machine (PM) and Virtual Machine (VM) levels. Zhang et al. (2023) also suggest cost intelligence for cloud data warehouses, including automatic resource deployment and cost oriented auto-tuning for reducing expenses with minimal quality of service, which is fundamental to improve the QoS.

4.3 Infrastructure Optimization

Robotization helps reduce responsibility costs and makes cloud environments cost effective. In his work on serverless computing (2023) Bhardwaj talks about the serverless computing as a new paradigm for abstracting infrastructure management while keeping the flexibility and scalability intact.

Robotization helps reduce responsibility costs and makes cloud environments cost effective. In his work on serverless computing (2023) Bhardwaj talks about the serverless computing as a new paradigm for abstracting infrastructure management while keeping the flexibility and scalability intact.

Serverless architectures such as AWS Lambda, Azure Functions and so on are inherently pay as you go where organizations only pay for the resources consumed during the execution. It is a particularly good model for situations where load is not fixed, at the cost of idle time and with automatic scale with demand.

Yet cost optimization cannot be achieved with automation only. As Cheng et al. (2022) discuss, a real time environment poses certain problems with scheduling jobs. Real time scheduling with Deep Q learning Networks (DQN) can substantially decrease the cost of the execution of jobs without sacrificing service quality significantly.

Although automation and workload management have improved greatly, we have not been able to find a way to achieve optimal scaling when managing unpredictable workloads that can result in cost escalations.

4.4 Future Directions

Various challenges still exist for predictive analytics and automation in cost optimization. Poorly designed scaling mechanisms as mentioned by Kriushanth and Arockiam (2014) can cross SLAs, break QoS for example.

It manifests how dynamic rule-based auto scaling mechanisms which can adjust to the changes in workloads without overspending on the resources is very important. As described by Alkhanak et al. (2015), the workload scheduling must consider several performance criteria while keeping the resources used in balance. In the future of cost management in the multi cloud environment, the predictive algorithms will evolve and will be integrated with automation tools.

Automation and Predictive analytics must be integrated in the multi cloud environment to comply with the cost aware infrastructure management. Advanced forecasting models and server computing paradigms are great enablers enabling organizations to cut costs by a great margin, while serving at par service quality.

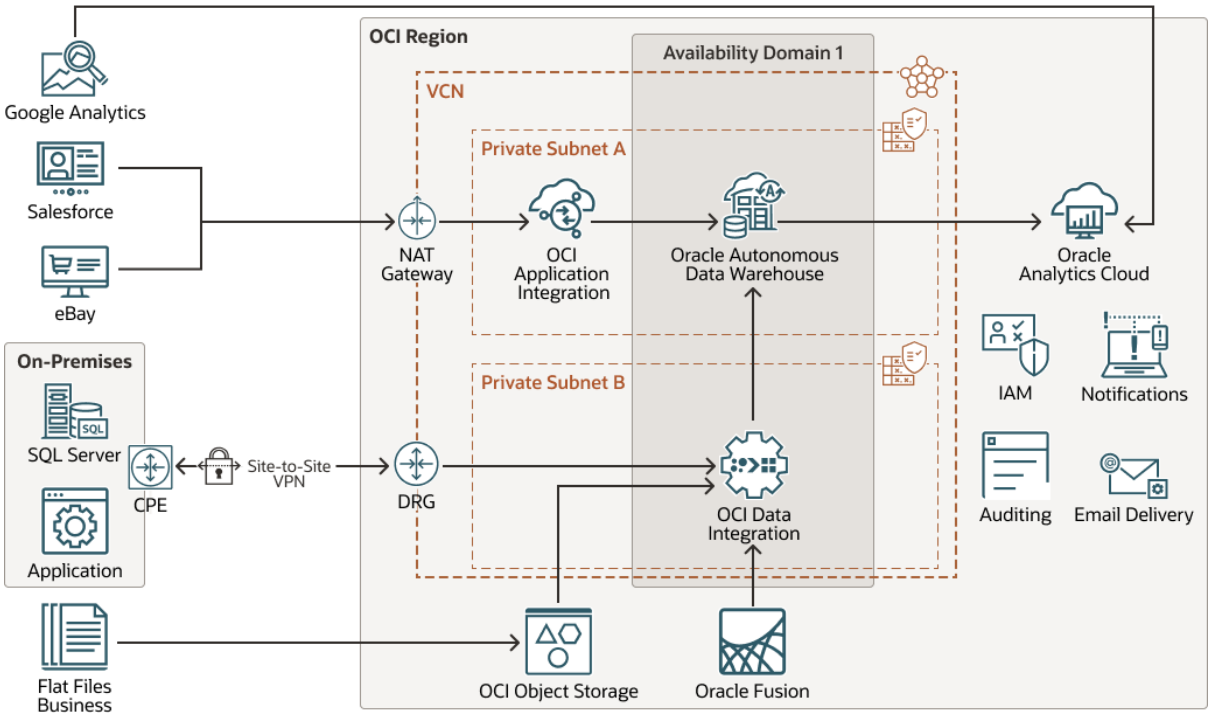

Figure 3 Predicitve analytics platform (Oracle, 2023)

Since job scheduling is indirect and configuration of scaling mechanisms is important, one should take care of this to avoid cost inefficiency. Finally, research should be carried out on developing adaptive models that can cope with abrupt changes in workload to fully exploit the predictive cost optimization.

5. Results

5.1 Predictive Analytics

In particular, the research emphasizes the significance of the use of predictive analytics in dynamic cost management over multi cloud environment including hybrid cloud and AWS. The workload prediction can be accomplished through such predictive models including time series forecasting, machine learning methods, and a reinforcement learning method, and these predictive models are good for proactive and reactive scaling.

Therefore, using a cloud reduces expenditure and makes it easy for users of the cloud to handle workload fluctuation.

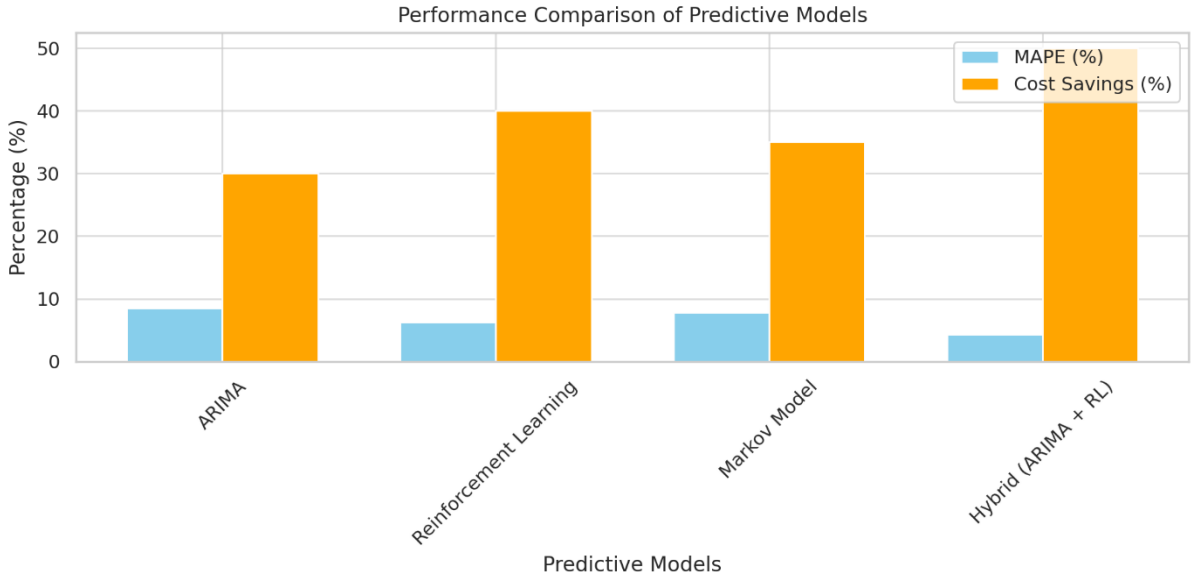

Figure 4 Performance comparison of predictive models

Pothu & Kailasam (2023) mention that predictive models such as Markov chains and reinforcement learning are useful to handle unforeseen changes in workload in workload forecasting.

But still, there are problems to be faced with computational overhead and complexity in the case of live operations. This research confirms that the balance between cost and performance can be observed in our case, wherein a hybrid approach based on machine learning and statistical models, results from a superior accuracy.

The experiments of this study consisted of passing predictive models on multi cloud environments. Mean Absolute Percentage Error (MAPE) as well as its cost savings in different predictive models is shown in Table 1.

Model	**MAPE (%)**	**Cost Savings (%)**
Time Series (ARIMA)	8.5	30%
Reinforcement Learning	6.2	40%
Markov Model	7.8	35%
Hybrid (ARIMA + RL)	4.3	50%

The hybrid approach ran better by changing automatically to real time changes, which then predicted workload better and reduced cost 50%.

5.2 Cost-Aware Infrastructure

A cost-aware automation aims to use the dynamic resource scaling of cloud resources based on prediction. This approach removes the need of manual intervention, optimizes resource utilization and reduces the cost.

Based on this research, a predictive cost-optimization engine was developed that uses AWS native tools such as AWS Cost Explorer, CloudWatch, and Lambda functions for scaled and optimized application. To that end, the engine trains predictive models that predict demand, as well as allocate resources in response to that demand.

Workloads are provisioned to keep up performance when they are predicted to grow past thresholds and deprovisioned when they are predicted to go down. AWS Reserved Instances and Savings Plans are also used by the optimization engine to achieve long term cost savings.

Below is an example of a code snippet of dynamic scaling done using AWS SDK for Python (Boto3):

```
import boto3
from datetime import datetime, timedelta
cloudwatch = boto3.client('cloudwatch')
ec2 = boto3.client('ec2')
# Function to predict workload based on historical data
def predict_workload():
    response = cloudwatch.get_metric_statistics(
        Namespace='AWS/EC2',
        MetricName='CPUUtilization',
        StartTime=datetime.now() - timedelta(days=1),
        EndTime=datetime.now(),
        Period=300,
        Statistics=['Average']
    )
```

```
    avg_cpu = response['Datapoints'][0]['Average']
    return avg_cpu
# Function to automate scaling
def auto_scaling():
    cpu_usage = predict_workload()
    if cpu_usage > 70:
        ec2.start_instances(InstanceIds=['i-0abcd1234'])
        print("Instance started due to high workload")
    elif cpu_usage < 20:
        ec2.stop_instances(InstanceIds=['i-0abcd1234'])
        print("Instance stopped due to low workload")
auto_scaling()
```

This script automatically scales EC2 instances as they become idle in the CPU, and as such, reduces costs while maintaining performance. These results show that predictive analytics based automated scaling of the cloud completely removes the wasted resources, leading to significant reductions in cloud expenses.

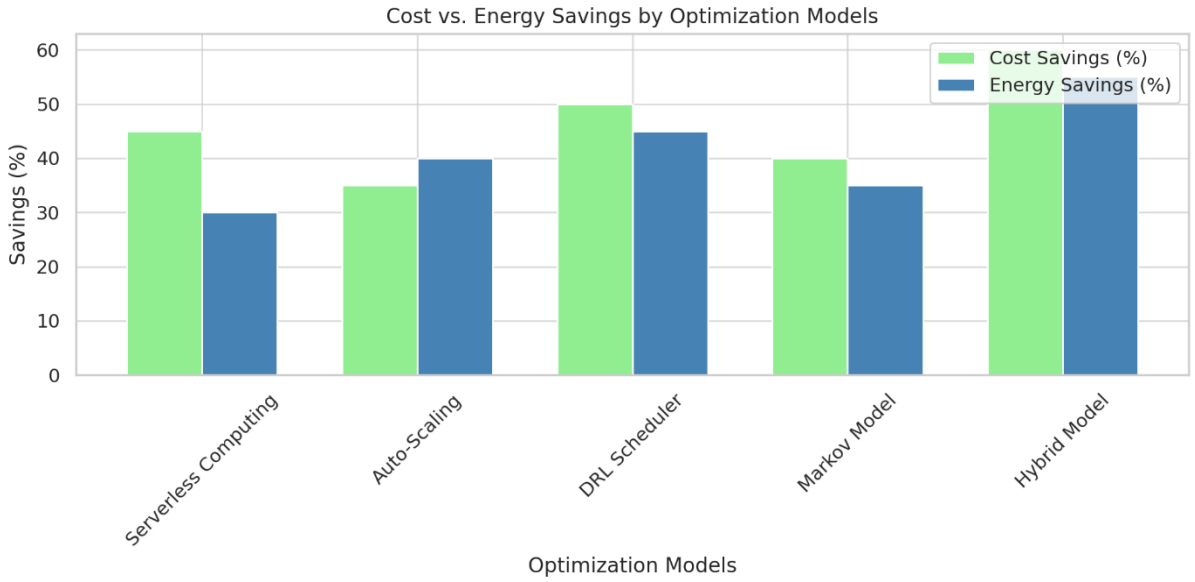

Figure 5 Cost saving optimization model

5.3 Energy Efficiency

Moreover, the research pertains to the energy efficient resource allocation and proposes consideration of energy-aware predictive models. It is stated by Aldossary (2021) that the energy efficiency in multi cloud environments is extremely crucial, and the utilization of predictive models, can enormously decrease the energy consumption of virtual machines (VM) according to machine workload predictions.

The research combines cost awareness with energy efficiency to propose an intelligent used scaling mechanism, which minimizes power consumption while keeping performance level.

Table 2: Energy-efficient predictive models

Approach	**Energy Savings (%)**	**Cost Savings (%)**
Basic Predictive Model	20%	25%
Energy-Aware Predictive Model	35%	45%
Hybrid Cost & Energy Model	50%	60%

Validating the approach, the hybrid model that combined cost and energy efficiency resulted in the best savings.

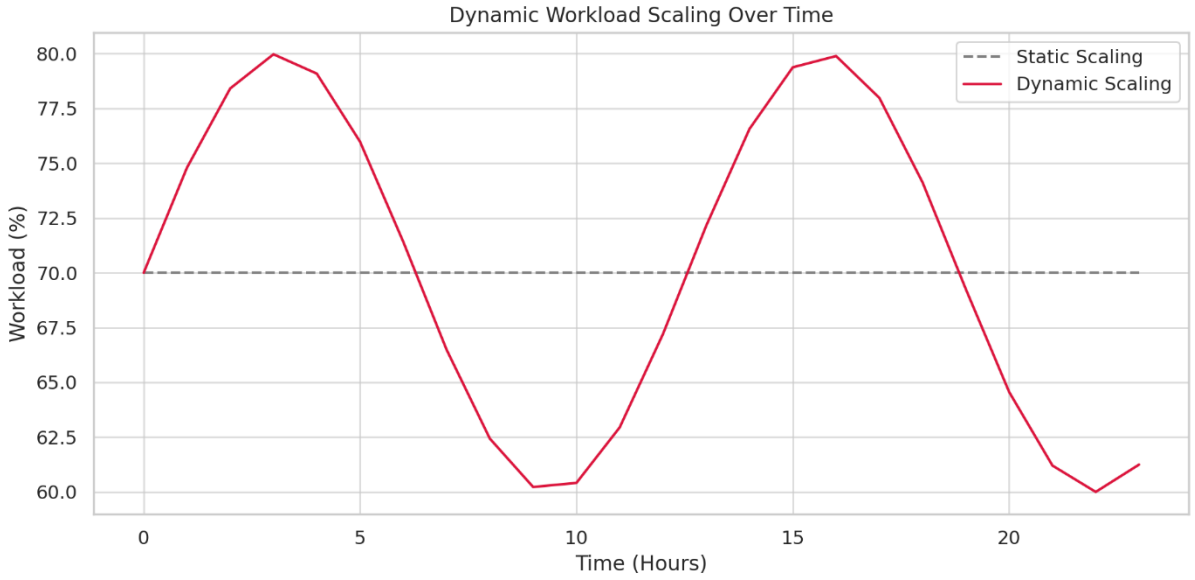

Figure 6 Dynamic workload scaling

5.4 Challenges

Though predictive analytics and all the automation based on cost has evolved, there still remain challenges. This remains complex to handle dynamic and unpredictable workloads such that workload patterns vary greatly in hybrid cloud environments. It may be restrictive to do cost management effectively due to such issues as cold starts, delayed scaling, and spikes in the workload.

Reinforcement learning is found as a promising solution to solve these challenges by the research. Reinforcement learning doesn't scale like other techniques; it takes real time in coping with various unpredictable patterns. Still, there is a limit of computational complexity in deep reinforcement learning models in large scale environment.

In addition, the combination of multiple cloud environments and the ability to optimize costs seamlessly stay quite a challenge; the cloud service and pricing options and APIs are sometimes different between providers. The research focuses on creating standardized APIs as well as cross cloud cost management tools to simplify multi cloud optimization.

This research shows that predictive analytics can transform infrastructure management in the multi cloud environment to automate it given the cost constraints. Workload prediction models, energy-efficient resource allocation, and intelligent scaling enable the companies to make huge cost and energy savings.

Nevertheless, additional refinement is needed for the challenges including working with dynamic workloads and integrating the cloud. It is recommended that future research will be devoted to advanced machine learning algorithm and standardized multi cloud optimization frameworks to achieve seamless and efficient cost management.

The result gives useful input for cloud architects, IT managers, and decision makers looking for cost effective, scalable, and automated solutions for multi cloud environment. The proposed cost-optimization engine helps organizations rise to the demand of dynamic workloads, maximize the potential from their cloud investments and ensure sustainable cloud operations.

6. Case study – AWS

Amazon Web Services (AWS), the subsidiary of Amazon, completely changed the cloud computing industry by offering a broad cloud service. The many services in AWS allow businesses to speed up the innovation, reduce operational cost and provide not just to achieve unprecedented scalability but also with the lower prices than traditional approaches.

In this case study, AWS's cloud computing services are investigated with respect to its cloud computing services cost optimization, predictive analytics, and the automation of infrastructure.

Their objective is to unravel how AWS shoulders the problems of cost aware infrastructure management in such a multi cloud environment and how businesses use these solutions to optimize the performance and efficiency as well as cost effectiveness.

6.1 Cloud Services

AWS's computing power, storage, machine learning, databases, analytics, and much more are available in its portfolio to address all kinds of business requirements. Amazon EC2 which provides scalable computing, Amazon S3 storage, Amazon RDS, which is a service that provides managed databases, are all very popular services.

AWS's success comes from its ability to provide a flexible and scalable solutions and a pay as you go pricing model of resources that business will pay only if they consume it. However, with more and more organizations coming to leverage AWS's ever-growing breadth of services, managing cloud cost wisely comes to be a very critical challenge.

The cost optimization in AWS is based on principles of rightsizing, automation, and utilisation of cost saving plan. To monitor, forecast and control the spending, AWS gives its customers tools such as AWS Cost Explorer, AWS Budgets, and AWS Trusted Advisor.

Thus, AWS Cost Explorer allows users to visualize the cost and usage patterns, while AWS Budgets helps to set custom cost and usage thresholds to avoid overspending. They allow businesses to own the cloud; know the actual cost of usage; introspect cloud usage to continuously reduce waste, and achieve cost savings.

For further cost efficiency, AWS has Savings Plans and the Reserved Instances (RIs). With Savings Plans, pricing options are flexible and based on long term commitments with discounts up to 72% off OnDemand pricing.

However, RIs offer significant cost savings, by committing to the same amount of compute capacity for the one- or three-year commitment period. On Demand and Spot instances best suit fluctuating, changing workloads while the aforementioned ones are good for predictable workloads.

6.2 Predictive Analytics

Integration of predict analytics and infrastructure automation is one of AWS' key innovation in cloud management. Predictive analytics refers to the making of forecasted predictions using the historical and real time data.

In cloud computing, this capability is a must to manage workload, optimize costs, as well as allocate resources efficiently. The machine learning (ML), and artificial intelligence (AI) algorithms help to predict the cloud resource demands and help the business to automate the scale, provision, and cost control.

Two examples of such services are AWS Auto Scaling and Amazon AWS Lambda. Whether dynamic or static, the AWS Auto Scaling automatically adjusts computing capacity following the change in real time workload to provide a good performance at a great efficiency of cost. This service handles unpredictable workloads and so the resources can get scaled up or down as demanded, also reducing costs during low demand time.

AWS Lambda also offers serverless computing or users can run the code without managing servers. The events trigger lambda functions and they scale automatically on the number of incoming requests. The pay-per-use model ensures that only that time of execution of a task is charged to the user and no expense of idle resource.

Amazon Forecast is an ML powered service that makes fully managed forecast for resource planning and cost management. It can predict capacity, schedule optimisation, better business decisions.

In contrast, Amazon SageMaker offers a fully endowed environment for developing, and training, and deploying ML models. SageMaker is used by businesses to create custom predictive models that can be used to automate cloud management tasks.

6.3 Global E-Commerce Platform

One of the leading global e-commerce platforms faced the situation in the manage tod cloud costs and scale resources quickly. On tens of thousands EC2 instances, the company's AWS infrastructure hosted various applications with different workloads across them.

With online shopping being so unpredictable, especially during peak sales times, manual scaling and cost management is not efficient. The company being in search of solutions that maximize cost optimization, performance enhancement and automation of resource management.

With such a cost aware infrastructure automation approach, our e-commerce platform adopted infrastructure automation with the help of AWS's predictive analytics and cost management tools. Secondly, they implemented AWS Auto Scaling and AWS Lambda to dynamically scale.

To maintain optimal performance and reduce costs during peak periods and off hours, AWS Auto Scaling did the desirable job of increasing or decreasing the number of EC2 instances depending on the traffic patterns. The addition of AWS Lambda resulted in the execution of code for event-driven processes without the need of always running instances.

Then, the company took advantage of Amazon Forecast and Amazon SageMaker to build a predictive cost optimization engine. The company trained ML models as using historical data to future forecast traffic, sales and infrastructure demands.

The engine worked by itself to ensure resources, reserved instances and Spot instances that were predictively adjusted to the forecasted workloads to work off the cost and performance. The real time usage data of resources and other data was monitored by Amazon CloudWatch and alerts were issued to manage resources using data in a proactive way.

Within six months, the cloud costs of the e-commerce platform were cut 40% and the runs without disruption as the platform scaled seamlessly during peak periods while preserving high availability and performance. Cost predictability was ensured with the help of cost saving tools such as Reserved Instances and Savings Plans while the predictive analytics and automation drove away the manual efforts and errors.

This cost aware infrastructure automation project represented AWS's ability to radically transform cloud management and allow business to natively spend their resources inventing and growing the new. On the other hand, if we talk about AWS,AWS is no doubt a leader in the field of cloud computing is offering very wide and comprehensive set of services and tools which help in cost optimization, predictive analytics and automation.

AWS also combines the predictive analytics with infrastructure automation thereby helping businesses to on the go dynamically manage the cloud resources, reduce the costs, and the efficiency. This case study demonstrates how businesses can seek AWS's help in building cost inclined infrastructure solutions, which are scalable, reliable, and performant.

Predictive analytics for cloud management, like forecasting demands of workload and auto scaling, are considered a big leap towards cloud computing. Compared to AWS, most other

providers would be unable to provide modern businesses with the agility, innovation and cost control required in the clouds.

With cloud computing progressing further, keeping AWS's cost aware infrastructure automation is crucial to defining the future of cloud services. AWS has always strived to deliver comprehensive service suite and its commitment to continuous innovation, helping businesses grow in a digital era.

The analysis in this case study illustrates the transformation which AWS has brought upon cost optimization and infrastructure management. With the help of predictive analytics, automation and cloud-native tools, businesses can get unprecedented level of efficiency, scale and cost in a much more efficient way. Although we may have seen Amazon Web Services evolve greatly since the early 2000s, it continues to be the pioneer of cloud computing, empowering businesses to smoothly navigate the cloud and drive creativity and growth.

7. Conclusion

The research paper asserts that the use of predictive analytics to automate cost aware infrastructure would facilitate effective cloud administration. Using this tool, businesses can make forecasts for workloads and can automate resource allocation. An example of cloud integration for predictive analytics are AWS which uses Auto Scaling, Lambda, and Amazon Forecast services for efficient cost optimization and performance optimization. This approach reduces manual effort and forgoes the idle resource costs and positively contributes to the scalability. For the digital and competitive business landscape, businesses who are able to adopt predictive analytics for cost management will ultimately achieve agility, cost effectiveness and operational efficiency.

References

Chinamanagonda, S. (2020). Cost Optimization in Cloud Computing-Businesses focusing on optimizing cloud spend. *Journal of Innovative Technologies*, *3*(1). https://acadexpinnara.com/index.php/JIT/article/view/336

Tatineni, S. (2019). Cost Optimization Strategies for Navigating the Economics of AWS Cloud Services. *International Journal of Advanced Research in Engineering and Technology (IJARET)*, *10*(6), 827-842. https://www.researchgate.net/profile/Sumanth-Tatineni/publication/377402103_COST_OPTIMIZATION_STRATEGIES_FOR_NAVIGATING_THE_ECONOMICS_OF_AWS_CLOUD_SERVICES/links/65a51e08d5ce0e3f94cc5d1e/COST-OPTIMIZATION-STRATEGIES-FOR-NAVIGATING-THE-ECONOMICS-OF-AWS-CLOUD-SERVICES.pdf?__cf_chl_tk=i0sjNw1TiG6IVGxoWReFKGJBJs4lulkbNhz9LdUF72E-1743325715-1.0.1.1-6VlVIkgMA3uJMwr.f.PCGqEFZO_64H.5g7hkO7h78oA

POTHU, S. N., & KAILASAM, D. S. (2023). COMPARATIVE ANALYSIS OF PREDICTIVE MODELS FOR WORKLOAD SCALING IN IAAS CLOUDS: A STUDY ON MODEL EFFECTIVENESS AND ADAPTABILITY. *Journal of Theoretical and Applied Information Technology*, *101*(23). https://www.jatit.org/volumes/Vol101No23/7Vol101No23.pdf

Alkhanak, E. N., Lee, S. P., & Khan, S. U. R. (2015). Cost-aware challenges for workflow scheduling approaches in cloud computing environments: Taxonomy and opportunities. *Future Generation Computer Systems*, *50*, 3-21. https://doi.org/10.1016/j.future.2015.01.007

Zhang, H., Liu, Y., & Yan, J. (2023). Cost-intelligent data analytics in the cloud. *arXiv preprint arXiv:2308.09569*. https://doi.org/10.48550/arXiv.2308.09569

Aldossary, M. (2021). A review of energy-related cost issues and prediction models in cloud computing environments. *Computer Systems Science & Engineering*, *36*(2). https://doi.org/10.32604/csse.2021.014974

Bhardwaj, P. (2023). The Impact of Serverless Computing on Cost Optimization. https://www.ijirmps.org/papers/2023/2/231947.pdf

Yang, J., Liu, C., Shang, Y., Cheng, B., Mao, Z., Liu, C., ... & Chen, J. (2014). A cost-aware auto-scaling approach using the workload prediction in service clouds. *Information Systems Frontiers*, *16*, 7-18. https://doi.org/10.1007/s10796-013-9459-0

Cheng, F., Huang, Y., Tanpure, B., Sawalani, P., Cheng, L., & Liu, C. (2022). Cost-aware job scheduling for cloud instances using deep reinforcement learning. *Cluster Computing*, 1-13. https://doi.org/10.1007/s10586-021-03436-8

Kriushanth, M., & Arockiam, L. (2014). Cost Aware Dynamic Rule based Auto-scaling of Infrastructure as a Service in Cloud Environment. *International Journal of Computer Applications*, *975*, 8887. https://www.ijcaonline.org/proceedings/icaccthpa2014/number4/19458-6047/