Interactive dashboard for Cost optimization of cloud resources using machine learning

A Project Report submitted by

Poornima H R

in partial fulfillment of the requirements for the award of the degree of

M.Tech.

Indian Institute of Technology Jodhpur
Name of the Department

January 2023

Declaration

I hereby declare that the work presented in this Project Report titled Interactive dashboard for Cost optimization of cloud resources using machine learning - M.Tech. submitted to the Indian Institute of Technology Jodhpur in partial fulfillment of the requirements for the award of the degree of M.Tech., is a bonafide record of the research work carried out under the supervision of Professor Dr. Sumit Karla. The contents of this Project Report in full or in parts, have not been submitted to, and will not be submitted by me to, any other Institute or University in India or abroad for the award of any degree or diploma.

Signature
Poornima H R
M21AI564

Certificate

This is to certify that the Project Report titled Interactive dashboard for Cost optimization of cloud resources using machine learning, submitted by Poornima H R (M21AI564) to the Indian Institute of Technology Jodhpur for the award of the degree of M.Tech., is a bonafide record of the research work done by her under my supervision.To the best of my knowledge, the contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Signature
Dr. Sumit Karla

Abstract

The project's goal is to create an interactive dashboard that will provide recommendations for cloud SME's to optimize costs of public cloud (preferably Azure or AWS) IAAS services using various AI and ML logic. Here, the users (ex., cloud SME's or cloud solution architects) can upload the cloud service usage data into an interactive dashboard and get predictions to lower the cost of cloud services based on the data they upload and visualize them effectively.

Contents:

1. Introduction and Background:

Providing an interactive dashboard to predict demand for computing resources and provide suggestions for cloud cost optimization is a vital task since it allows the optimized management of resources. Developing a platform to get insights about computing cloud environment services, viz., storage, network, and computing resources, is essential for knowing the cost-optimized solutions for utilizing the resources offered by different cloud service providers (AWS, Azure, GCP, etc.).

Here we consider the usage of virtual machines and databases. Our solution can reduce cloud resource usage costs by forecasting demand for various resources (for example, CPU, IOPS, memory, and storage) and then adjusting cloud components accordingly. Prediction is accomplished through the application of various machine learning techniques. This system allows the users (ex., cloud SME's or cloud solution architects) to upload the cloud service usage data into an interactive dashboard and get predictions to lower the cost of cloud services based on the data they upload and visualize them effectively. This will be achieved by building a small system website that uses micro-service architecture and deploying those services on the cloud using a different deployment architecture.

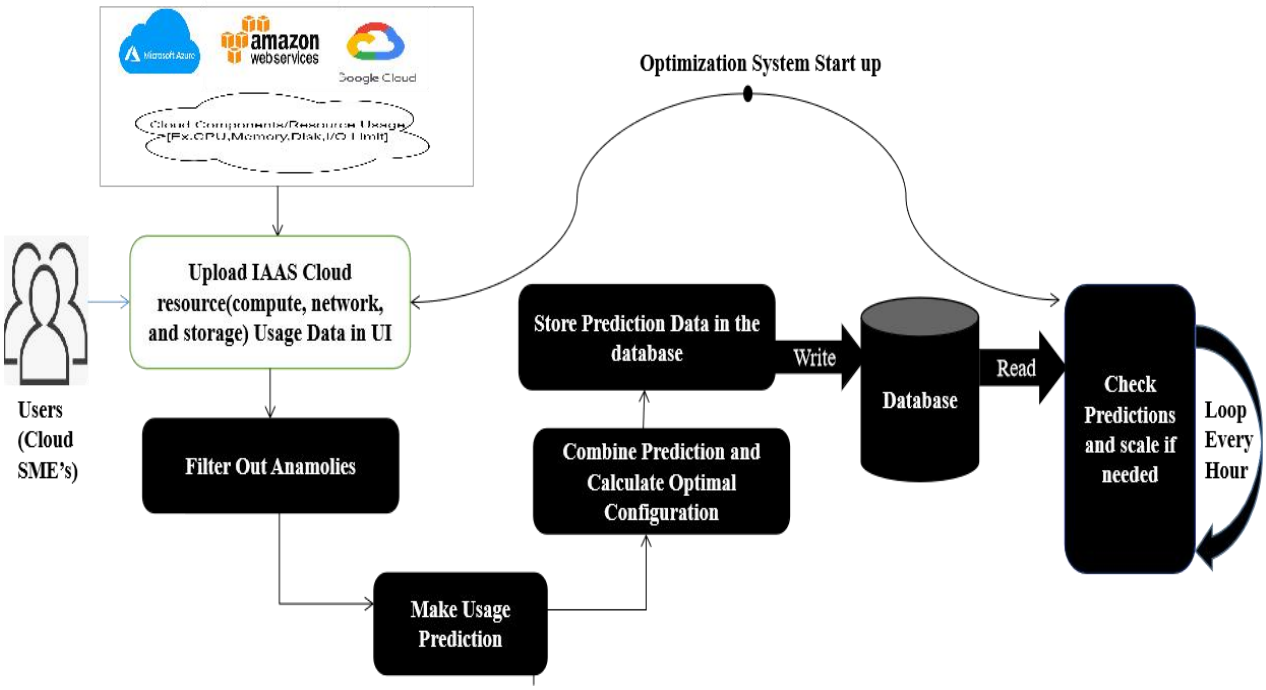


Fig 1.1. Resource Usage and cost optimization setup overview

2. Literature Survey:

The author [1] provides us with an understanding of how to achieve a cost-optimal cloud resource configuration using anomaly detection, machine learning, and particle swarm optimization techniques, which are essential for the fast-growing cloud era. This article also provides knowledge about how the resource usage can be optimized by using anomaly detection technique. Even though the author of the article have tested this in Microsoft Azure cloud environment , this approach can be used in other clouds as well such as AWS, GCP etc.,

The primary prediction module algorithm used in implementation is mentioned as below.

```
for all component of the monitoredComponents do
    for all resources of the component do
        data = GetHistoryData(resource);
        data = AnamolyFilter(data);
        data = MedianFilter(data);
        WriteToDB(data);
        predictionData = ReadFromDB(configureWindow);
        predictionResult = predictionUsageML(predictionData)
predictions.Add(predictionResult);
    end for
end for
pricing = GetPricingPlan();
configuration = CalculateConfiguration(prediction,pricing);
WriteToDB(configuration);
```

With the help of this algorithm, we'll be able to make predictions about the optimal configuration of cloud resources hosted in a cloud environment and store them in a database to provide an effective pricing plan and calculate the optimal resource configuration.By considering the threshold of resources (i.e., CPU, memory, disk, etc.) and the types of configurations done in components by the cloud providers or the resource users, the author will be finding a set of configurations that will meet the given constraints and provide a cost-efficient solution.The solution proposed by the author in the cloud computing field of research is effective, helps in optimization of the resource at the allocation level, and can

effectively be used in cloud-based systems that use scalable resources (i.e., IaaS, PaaS, or SaaS).

The article[2] describes the development of a hybrid model for the efficient utilization of cloud resources. Because cloud computing has significant **challenges with resource allocation and task scheduling,** the author initiated the research and proposed a hybrid modelling technique to maximize the utilization of cloud computing resources **using machine learning and particle swarm optimization techniques** for efficient cloud resource provisioning. The proposed approach can be used to optimize task scheduling in a heterogeneous cloud computing environment. The **particle swarm optimization** (PSO) **algorithm** for task scheduling is developed and used for this purpose. The proposed model would be run on **CloudSim**, and it would make use of the PSO-ANN algorithm to provide an **optimal solution for job scheduling.**According to the research article, cloud resource provisioning is accomplished through the use of hybrid models (PSO and ANN). By observing the outputs obtained from cloud simulation, the given approach from the author of the article provides a quicker rate of throughput and a shorter Makespan while simultaneously reducing the amount of delay and helping in optimal task scheduling and resource allocation.

In Article [3], author proposes a solution for dynamic optimal resource allocation for network functions.Virtualization (NFV) components by using a method that combines the Markov Decision Process and Bayesian learning approach to dynamically allocate cloud computing resource for NFV components. While MDP helps to dynamically allocate NFV components to cloud resources, and applying machine learning methods to historical data to predict future resource reliability helps enhance the performance of NFV-oriented cloud resource management. Conducted evaluation proves that the proposed method outperforms other greedy methods in overall cost.However, the author wishes to enhance research work in three directions. First, to study the impact of cost models on different resources and to analyze the method performance amongmultiple competitive service providers. Second, we plan to consider NFV component dependencies among multiple tenants to build a more comprehensive model. Third, we plan to study other optimization algorithms, such as the Ant Colony Optimization Algorithm, in combination with machine learning techniques to further optimise the NFV-oriented resource allocation problems.

The proposed approach in article [4] is capable of helping to improve the application architecture and supporting the software architect in identifying an adequate architectural solution

while keeping costs under control. However, future enhancement work will consider the optimization of mixed IaaS and PaaS as well as multi-cloud applications, also considering the availability metric among performance constraints.

The solution provided in article [5] is LTPS(Long-Term Prediction System) which automatically determines the relationship between optimized system load and its QoS parameters by performing a series of trials. Machine learning and anomaly detection techniques are used for predicting resource provisioning.The solution proposed by the author uses anomaly detection and machine learning in both the discovery and execution stages, which use a QoS ML model that predicts resource levels and scales optimized system components.We can see the comparison of costs with respect to different cloud environments after using optimization techniques (see above table).As a result, the proposed optimization algorithm for hybrid cloud work-flow scheduling with the two-staged approach, anomaly detection, and dedicated Integer-PSO algorithm, which also brings QoS-constrained optimization efficiency improvements and industry-grade readiness.

The paper[6] presents the HCOC (Hybrid Cloud Optimized Cost Scheduling) algorithm, which helps in deciding the best resources to request from a public cloud based on demand and on resource costs in a hybrid cloud environment.It also guides the users about which resources should be leased from the public cloud and aggregated to the private cloud to provide sufficient processing power to execute a work-flow within a given execution time.In this article, the author presents extensive experimental and simulation results, which show that the HCOC can reduce costs while achieving the desired execution.As per the article, HCOC is an algorithm used to speed up the execution of work-flows in a given amount of time with reduced costs when compared to the greedy approach. This algorithm serves as a cost optimization algorithm for work-flow scheduling in hybrid clouds.

In Article [7] author has prototyped MArk on AWS and showed that compared with the premier autoscaling ML platform SageMaker, MArk yields significant cost reduction (up to 7:8) while complying with the SLO requirements with even better latency performance.However MArk’s architecture also requires a centralized master machine to make provisioning decisions, but Centralised model is vulnerable to single point of failure and has scalability issue.So this model MArk’s master node only performs lightweight computations. The potential scalability and reliability problems can be easily addressed with mature industrial solutions.

3. System Design and Analysis:

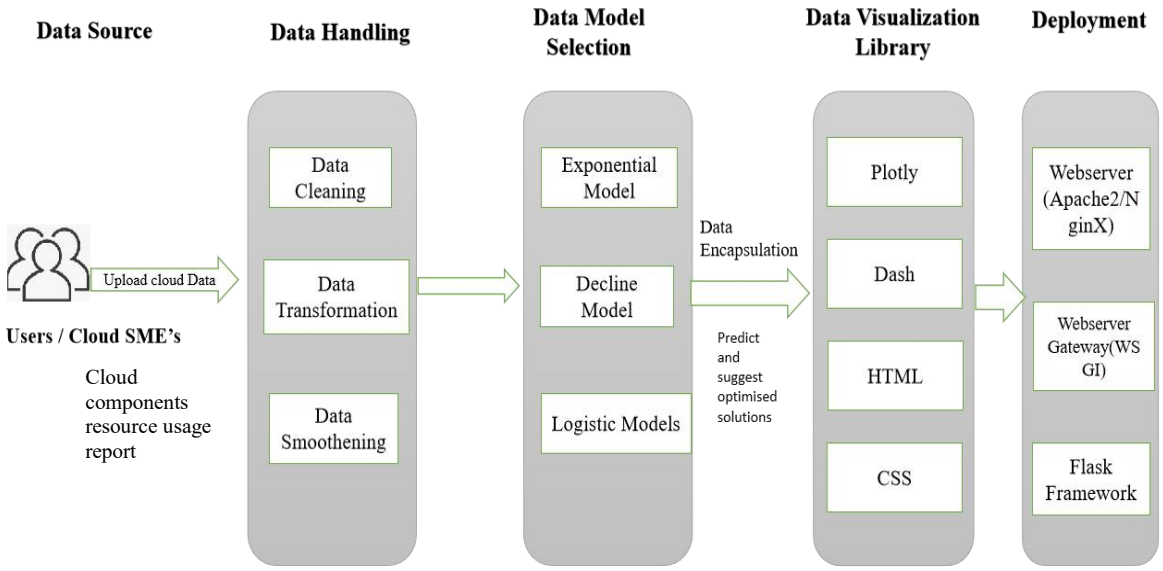


Fig.3.1. Architecture for creating an interactive dashboard for cost optimization of cloud resources using Machine Learning.

We're building a small system website which is using microservice architecture and deploy those services on cloud in a different deployment architecture.

Cloud service providers (viz., AWS, Azure, GCP, etc.) offer different services to components like VMs, DBs, and micro-services, and each component consists of different properties (i.e., compute power (CPU), random access memory size (RAM), disc capacity, input/output operations per second (IOPS)).

We need to utilize these historical cloud component usage reports and provide the process of scaling system components while taking the predicted usage level into account. In the process, we take into consideration the usage of virtual machines, application services, and databases from the perspective of a cloud solution architect in order to minimize the cost of cloud resource usage.

Our solution can reduce cloud resource usage costs by predicting demand for various resources (e.g., CPU, IOPS, memory, storage) in the cloud platform to handle various applications present in computing systems and then adjusting cloud components accordingly. Prediction is done with the use of machine learning techniques and AI algorithms, and these data can be visualized to get meaningful insights from the data in an interactive dashboard.

3.1 Machine learning Life cycle for Cloud Data set Processing:

- Learn from historical cloud data.
- Prepare a dashboard and visualize the data.
- Monitor the performance and try uploading different patterns of data.
- Train the machine using some correlation methods or different ML models.
- Consider the parameters and features that we're measuring, viz., performance, capacity, security, etc., and then find correlations between these for different use cases.
- Create a dashboard where SMEs can upload historical data and use it to visualize or gain insight into the values of selected parameters in the near future based on current data.
- We need to create a dashboard where SMEs can login to the dashboard and upload historical or current cloud resource usage reports and get insights about future data.
- Login > Upload Data > Model Selection > Predictions

  - **Login**
  - **Upload Data**: Different types of data may be part of the report.
  - Numerical data
  - bivariate ,multivariate data sets.
  - Categorical data sets.
  - Correlation data sets
  - **Select Model**: Based on the type of data, select an appropriate pre-coded model.
  - Classification Model
  - Regression Model
  - Exponential Growth and decline Model
  - Logistic growth Model

- Post uploading the data we can visualize the data statistically, Here we need to design an AI model to find a few Patterns.Example: Information about outliers(When these Exceptions or Outliers are occurring ??), Relationships between data , etc.,
- Then Claim the model for identifying these patterns.

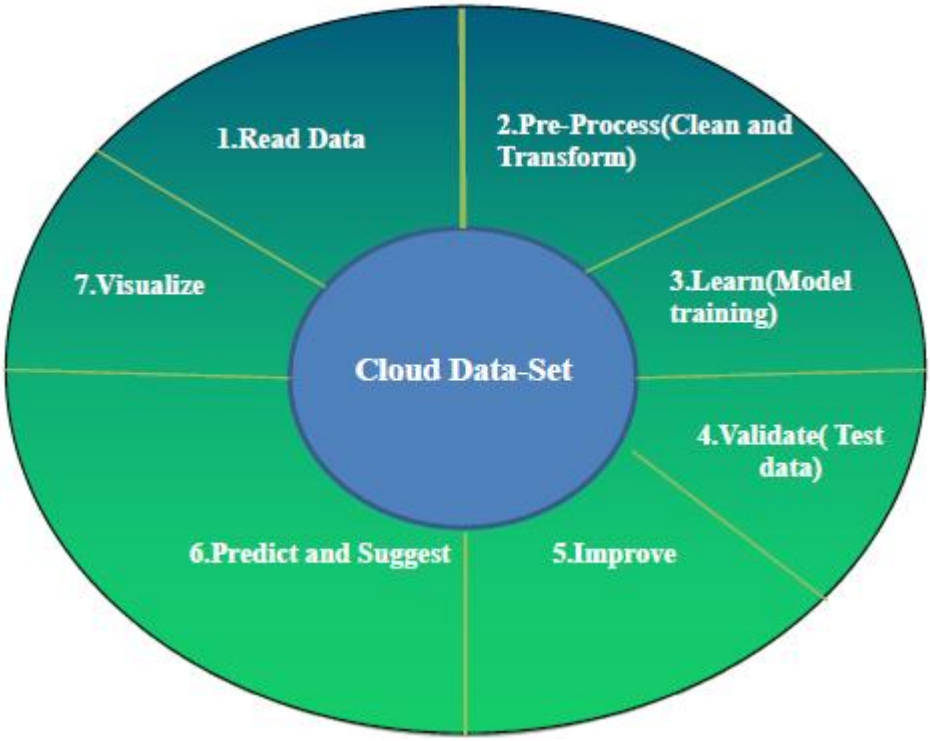
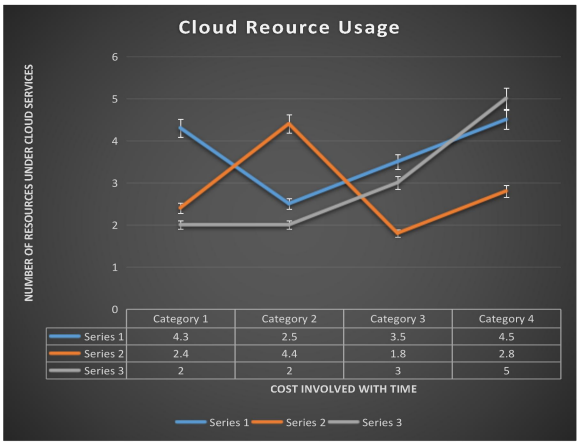


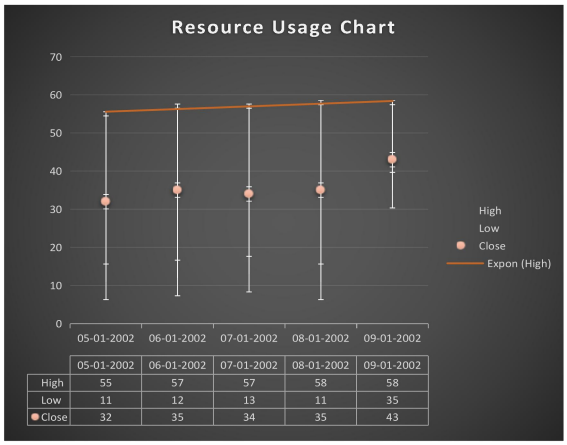
Fig.3.1.1. Machine learning Life cycle for Cloud Data set Processing

4. Data Analysis:

By seeing the training datasets, we extract the useful data by selecting different columns and passing them as parameters to our model, then storing the trained model using a file system. Then pass the test data and get results for cost optimization suggestions.For example, we can get reports in the dashboard to analyse CPU and memory utilisation of cloud components and determine which resources are underutilized, and with that information, we can recommend to the user that they resize the components to reduce the costs of those resources.



	Category 1	Category 2	Category 3	Category 4
Series 1	4.3	2.5	3.5	4.5
Series 2	2.4	4.4	1.8	2.8
Series 3	2	2	3	5



	05-01-2002	06-01-2002	07-01-2002	08-01-2002	09-01-2002
High	55	57	57	58	58
Low	11	12	13	11	35
Close	32	35	34	35	43

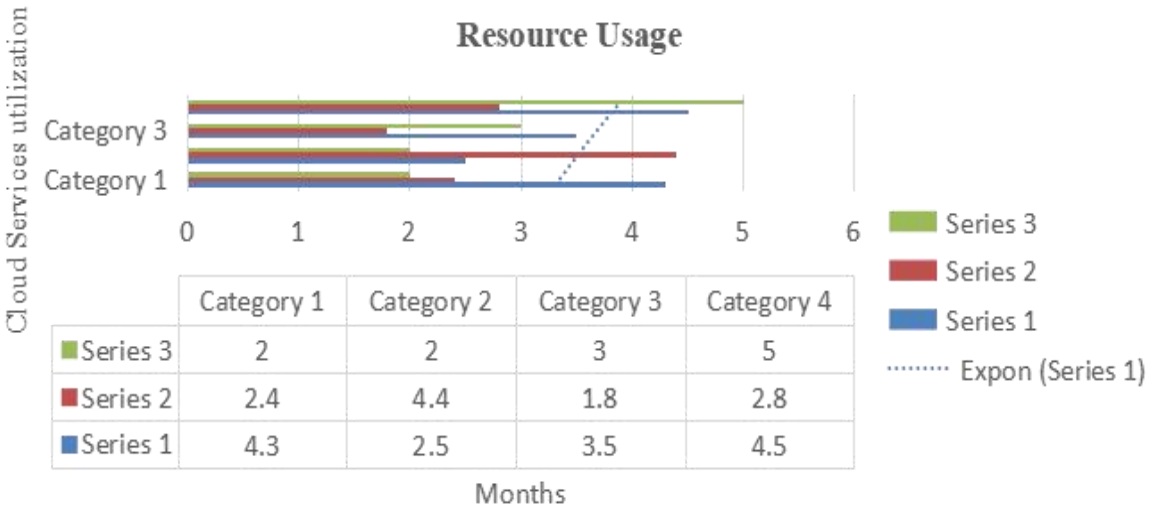


Fig.4.1. Cloud Services utilization Report

5. Requirement Specification:

Requirement specification includes both hardware and software requirements. The software requirement includes platform requirements, framework requirements, operating system supported and the web browser support.

Software Requirements	Python
	WSGI toolkit
	Jinja2 template engine
	Plotly,Dash,HTML,CSS
	Flask(Micro Web Framework)
Hardware Requirements	Processor: 64-bit, four-core, 2.4GHz or higher per core
	Disk Space Required: 5GB Recommended, 40GB minimum.
	RAM: 4GB RAM.
	OS: Windows
Public Cloud Platform	Azure/AWS

Table 5.1. Requirement Specification

6. Advantages/Uses :

➢ There is no open-source product or system to upload cloud service-related data and get predictions for having cost-optimized solutions for their cloud environment.

➢ Our product is going to provide an interface where the cloud experts can upload their data, have cost optimization to lower the cost of cloud services based on the data they upload, and effectively visualize those data.

➢ We're giving a pipeline to the non-programmer , where he can upload the data and extract useful data by selecting different columns and passing them as parameters to our Model and store the trained model using a file system.. Then pass the test data and show results.

AWS/Azure Dashboards	Proposed Interactive Dashboard
Doesn't have the option to upload files and get predictions for cost optimization of Resources.	Upload features for users will be available to get the necessary information.
Insights will be available only for existing/running resources data.	Insights for both old and newly added data will be provided.
Cost Optimizer or Advisor options will be given for used resources to get recommendations for cost optimization.	To develop and design new cloud architectures , solutions Architects can use the dashboard for optimal resource usage predictions.
Only the root user of an account can access the Billing and Cost Management console to understand more about the cost of an existing environment.	New Solution Architects or Cloud Experts can access the interactive dashboard without requiring root account access.

Table 6.1 Comparison between Existing Cloud Platform Reports and Our Interactive Dashboard

References :

[1] P.Osypanka and P. Nawrocki, "Resource Usage Cost Optimization in Cloud Computing Using Machine Learning," in IEEE Transactions on Cloud Computing, vol. 10, no. 3, pp. 2079-2089, 1 July-Sept. 2022, doi: 10.1109/TCC.2020.3015769.

[2] R. Atyam, R. Babu P, N. R. Nayak, U. Saikia, K. S. Ananda Kumar and A. S, "An Hybrid Modelling for Optimal Utilisation on Cloud Resources using Machine Learning Algorithm," 2022 International Conference on Electronics and Renewable Systems (ICEARS), 2022, pp. 1399-1403, doi: 10.1109/ICEARS53579.2022.9751806.

[3] R. Shi et al., "MDP and Machine Learning-Based Cost-Optimization of Dynamic Resource Allocation for Network Function Virtualization," 2015 IEEE International Conference on Services Computing, 2015, pp. 65-73, doi: 10.1109/SCC.2015.19

[4] Ciavotta, G. P. Gibilisco, D. Ardagna, E. D. Nitto, M. Lattuada and M. A. A. da Silva, "Architectural Design of Cloud Applications: A Performance-Aware Cost Minimization Approach," in IEEE Transactions on Cloud Computing, vol. 10, no. 3, pp. 1571-1591, 1 July-Sept. 2022, doi: 10.1109/TCC.2020.3015703.

[5] Nawrocki, P., Osypanka, P. Cloud Resource Demand Prediction using Machine Learning in the Context of QoS Parameters. J Grid Computing 19, 20 (2021). https://doi.org/10.1007/s10723-021-09561-3

[6] Bittencourt, L.F., Madeira, E.R.M. HCOC: a cost optimization algorithm for workflow scheduling in hybrid clouds. J Internet Serv Appl 2, 207–227 (2011). https://doi.org/10.1007/s13174-011-0032-0

[7] C.Zhang, M. Yu, W. Wang and F. Yan, "Enabling Cost-Effective, SLO-Aware Machine Learning Inference Serving on Public Cloud," in IEEE Transactions on Cloud Computing, vol. 10, no. 3, pp. 1765-1779, 1 July-Sept. 2022, doi: 10.1109/TCC.2020.3006751.

[8] https://www.geeksforgeeks.org/introduction-to-particle-swarm-optimizationpso/

[9] https://www.densify.com/resources/cloud-optimization

[10] https://www.sciencedirect.com/science/article/pii/S1877705812023259

[11] https://www.cloudability.com

[12] https://learn.microsoft.com/en-us/azure/cost-management-billing/costs/quick-acm-cost-analysis

[13] https://aws.amazon.com/aws-cost-management/aws-cost-explorer/

[14] https://aws.amazon.com/premiumsupport/technology/trusted-advisor/

[15] https://azure.microsoft.com/en-us/products/advisor/