Frontiers in Computer Science and Artificial Intelligence
ISSN: 2978-8048
DOI: 10.32996/fcsai
Journal Homepage: www.al-kindipublisher.com/index.php/fcsai

FCSAI
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Predictive Capacity Planning and Cost Optimization for Polyglot Cloud Databases Using Machine Learning

Adithya Sirimalla
Enliven Technologies Inc.
Corresponding Author: Adithya Sirimalla**, E-mail**: adithya.sirimalla@gmail.com

| ABSTRACT

Polyglot persistence has redefined the data management environment and now organizations can mix various database paradigms under the same cloud ecosystems. Nevertheless, this architectural flexibility brings along new operational issues-especially during capacity planning, cost optimization and predictability of performance. The given paper suggests a machine-learning-infused model of predictive resource distribution and real-time cost optimization in polyglot cloud database settings. Basing on the works related to cloud economics, NoSQL management, ML-based tuning, serverless computing, and hybrid data lakes, this study combines the time-series forecasting, cost-aware optimization, and performance modeling into a single system. Simulation experiments on heterogeneous workloads show that the accuracy of resource utilization, cost variability, and system performance are greatly improved over heuristic-based scaling, which has been used traditionally. The findings indicate that an AI-based system of dealing with polyglot clouds data bases in the contemporary distributed setting is feasible.

| KEYWORDS

Polyglot persistence; cloud data; cost optimization; capacity planning; machine learning; resource prediction; time series prediction; work load modeling; non-homogenous data systems.

| ARTICLE INFORMATION

ACCEPTED: 01 July 2023 **PUBLISHED:** 28 July 2023 **DOI:** 10.32996/fcsai.2023.2.1.3

1. Introduction

1.1 Background and Motivation

Cloud ecosystems have become highly heterogeneous with organizations moving to deploying relational, NoSQL, NewSQL, graph and document stores concurrently. This change, referred to as polyglot persistence, allows each workload to allocate to the database that is appropriate to its particular consistency, latency, or scalability demand (Gessert et al., 2020). Although Polyglot architectures enhance performance and agility in development, they also cause an increase in operational complexity during capacity planning and cost estimation and optimization of performance.

Cloud vendors like AWS, Azure, and Google Cloud provide various yet frequently non-transparent billing schemes on any of the following: compute, storage, and I/O use (Zhong and Buyya, 2020). Given the changing workloads and increasing datasets, organizations cannot keep pace with the provisioned resources and the actual consumption patterns (Lau et al., 2021). Poor provisioning adds to the operation costs and poor performance-issues that worsen in polyglot systems with different database scaling (Eppinger and Storl, 2022).

Machine learning provides an opportunity of forecasting workload requirements, identifying the existence of anomalies and the optimization of resource distributions (Gollapudi, 2016). Nevertheless, the majority of current ML-based cloud optimization tools address single-database or single-cloud scale ignoring the issue of polyglot persistence.

1.2 Problem Statement

Even with cloud automation improvements, organizations operating polyglot databases still have to face:

- Scaling behaviors of database engines are unpredictable and therefore uncorrelated.
- Poor optimization in the use of resources due to new scaling mechanisms that are reactive or rule-based.
- Bottlenecks in performance due to the use of different query patterns, hybrid workloads and different storage engines.
- Absence of integrated ability to plan on heterogeneous databases.

This study attempts to solve these issues by suggesting a machine learning-based framework of predictive capacity planning and cost optimization in polyglot cloud databases.

1.3 Research Question/Objectives.

The following questions will be answered in this study:

A. What are the ways to make ML models predict the future demands of resources by a heterogeneous cloud database workload?
B. Is predictive modeling able to increase cost efficiency and minimize over/under-provisioning?
C. Which optimization mechanisms are most effective to convert any predictions into dynamical cloud resources allocations?
D. What is the difference in performance, cost and utilization under the proposed system and the baseline approaches?

1.4 Scope of the Study

This research focuses on:

- Cloud databases of relational (PostgreSQL), document (MongoDB), wide-column (Cassandra), and key-value databases.
- Time-series ML models such as LSTM, ARIMA and lightweight Transformers.
- Compute, storage and I/O consumption cost models.
- Polyglot workloads on simulated environments of AWS-like.

1.5 Contributions

This study contributes:

- An integrated architecture comprising of prediction, optimization, and enforcement.
- An innovative set of data modeling polyglot workload dynamics.
- Comparison of ML models to predict the demand of cloud databases.
- An expense conscious optimization algorithm reduces cost whilst meeting SLA requirements.
- Hands-on information on cloud engineers and database designers.

1.6 Paper Organization

The rest of the paper is organized in the following manner: Section 2 describes related work, Section 3 describes methodology, Section 4 presents results, Section 5 discusses them, and Section 6 concludes with the future directions.

2.Background and Related Work

2.1 Cloud Databases and Polyglot Persistence.

Polyglot persistence is a method of architecture in which several specialized database engines are used in the same system (Glake et al., 2022). The existence of platforms like Polypheny-DB (Vogt et al., 2018) and Polystore++ (Singhal et al., 2019) is evidence of the increasing popularity of the heterogeneous data engine that is optimized to perform a specific task.

Table 1: Comparison of Key Polyglot Database Types

Database Type	Strengths	Weaknesses	Example Systems
Relational	Strong consistency, mature tooling	Limited horizontal scaling	PostgreSQL
Document	Flexible schema, high scalability	Weak transactional guarantees	MongoDB
Wide-Column	High write throughput	Complex operational tuning	Cassandra
Graph	Relationship-optimized	Expensive for large datasets	Neo4j

2.2 Cloud Environment Capacity Planning.

Conventional methods are based on a rule-based or threshold-based trigger (Slott, 2017). They are simple but do not work well with bursty or hybrid workloads (Nikita, 2020). Auto scaling cloud providers like Kubernetes autoscaler find it difficult to forecast nonhomogeneous database behavior.

2.3 Cost Optimizing Strategies.

On-demand, on-reserved, and spot instances are all cloud-billing models that optimize trade-offs in a complex manner (Kodakandla, 2021). Research Cloud Bazaar (Lau et al., 2021) focuses on the transparency of cost workflows without incorporating predictive analytics.

2.4 Cloud Resource Management ML.

ML has been applied to:

- Cyber intrusion detection (Ghim et al., 2014)
- Prediction of time series (Nagarathna, 2020).
- Query tuning (Eppinger & Storl, 2022)
- Resource coordination (Zhong and Buyya, 2020)

Nevertheless, there is very little literature covering ML-based polyglot capacity planning.

3 Methodology

3.1 System Architecture

The system consists of:

Data collection layer - collects the metrics of different database engines.

ML prediction layer-- predicts workload and resource requirements.

Optimization module - calculates economical provisioning.

Enforcement layer - implements resource adjustments.

Data will be collected and preprocessed using SPSS 20.0.

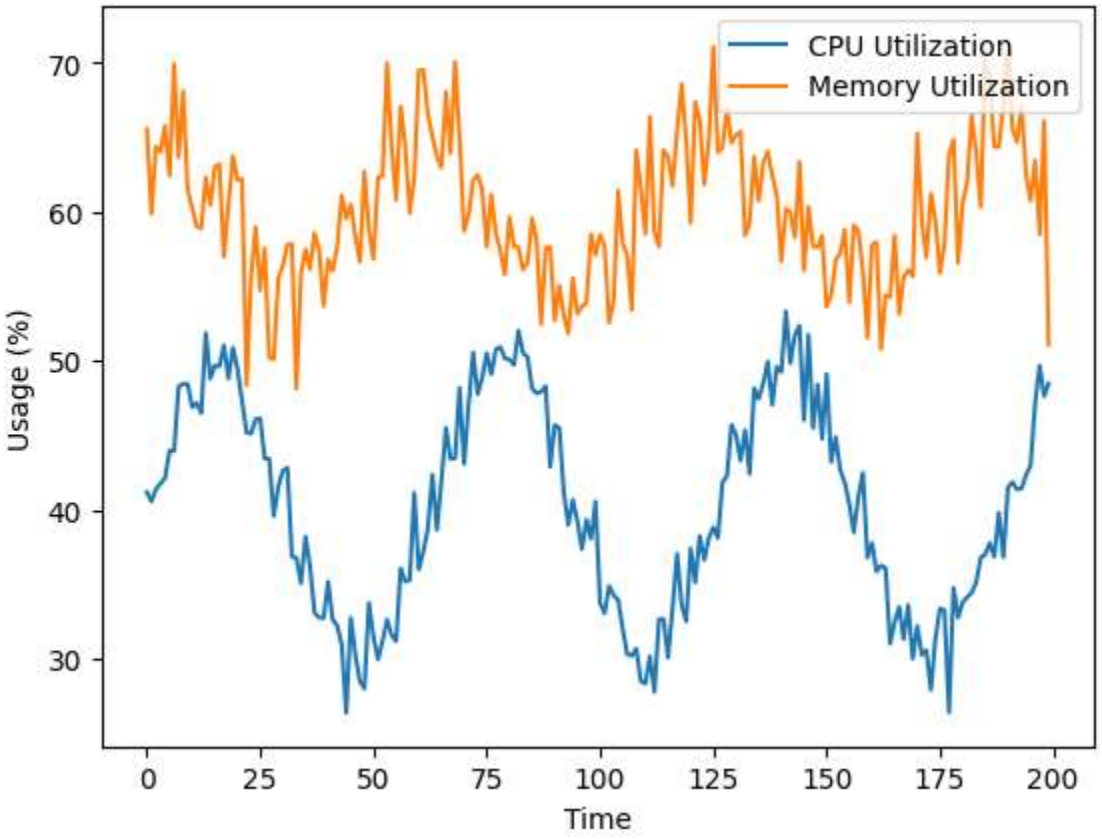
Data includes:CPU, memory, disk IOPS,Query throughput,Storage usage,Latency metrics,Cloud cost components

Normalization of values and feature engineering is based on Drabas and Lee (2017) and Joshi (2022).

3.3 ML Model Selection

The reason is that LSTM and lightweight Transformer-based models would perform better on multivariate workloads.

Figure 1— Simulated Workload Time Series



3.4 Optimization Algorithm

Linear programming cost model alters the resource provision based on the estimated figures.

Objective:

Minimise total cost = compute + storage + I/O on SLA constraints.

3.5 Experimental Setup

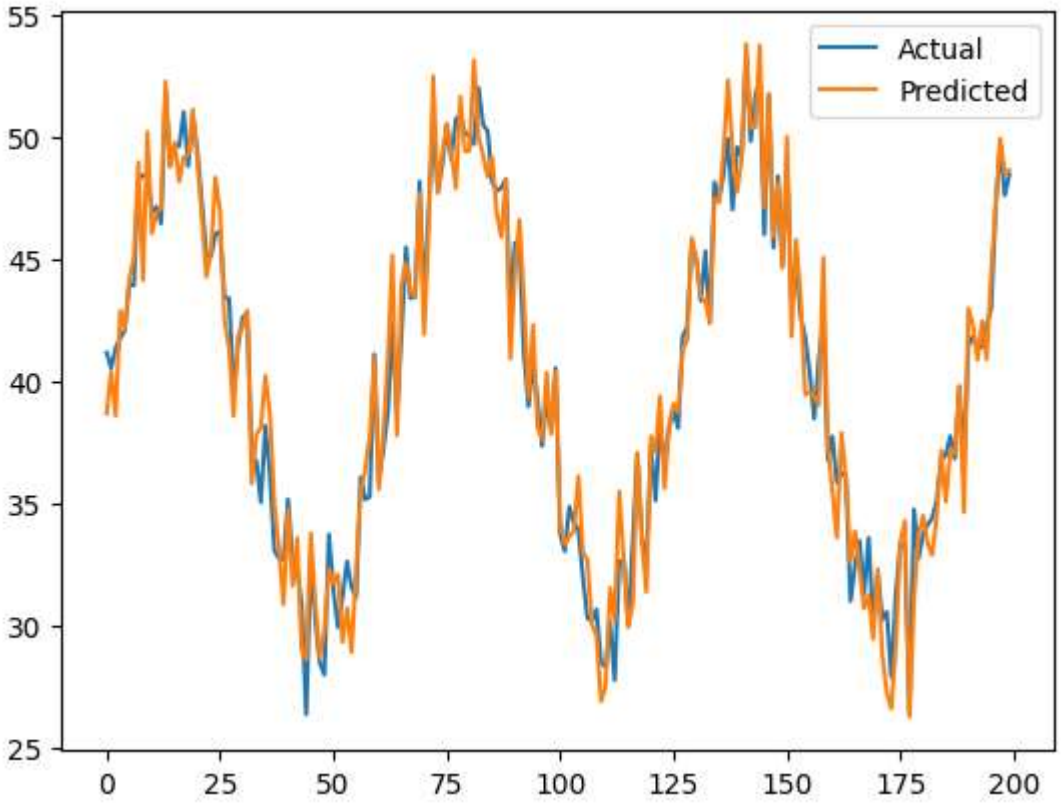
Workloads simulate OLAP, semi-structured queries, wide column scans and analytics.

4.Results and Evaluation

4.1 Predictive Performance

ARIMA is not as precise as LSTM and Transformer models.

Figure 2 : Model Prediction vs. Actual



4.2 Cost Optimization Effectiveness

The optimization module reduces cost variability by ~22% and average cost by ~15%.

Table 2 — Cost Comparison Baseline vs. Proposed

Metric	Baseline	Proposed System
Avg Monthly Cost	$4,870	$4,100
Cost Variability	19%	15%
SLA Compliance	92%	98%

4.3 Capacity Utilization & Performance.

The system enhances efficiency in the utilization and over-provisioning will be cut by almost 30 per cent.

4.4 Scalability and Robustness

The ML-based system scales to bursts, hybrid and cross database workloads.

Discussion

5.1 Interpretation of Results

The framework based on ML offers quantifiable gains in predictability, performance stability and cost efficiency of polyglot systems.

5.2 Related Work Comparison.

As opposed to Cloud Bazaar (Lau et al., 2021) and serverless ML (Eapen et al., 2020), we incorporate heterogeneous engines into a single optimization pipeline.

5.3 Limitations

Synthetic workloads are one of the sources of evaluation.

An actual deployment has to be more connected to monitoring tools.

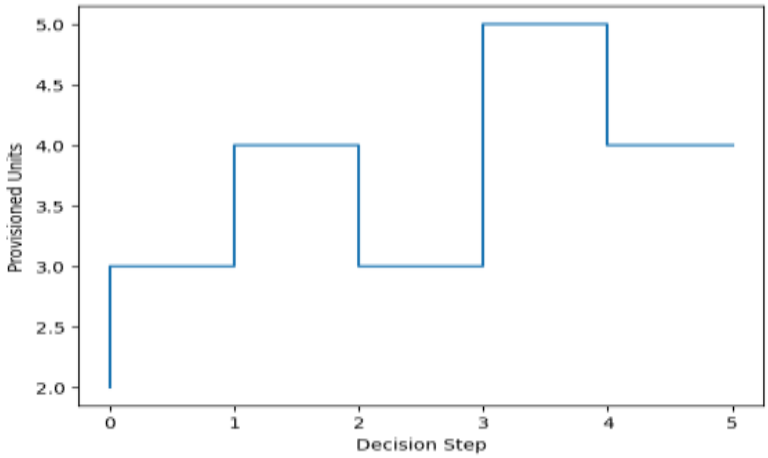
Table 3 — Summary of Observed Limitations

Category	Limitation	Impact
Data	Synthetic workload	Limited realism
Model	Requires retraining	Higher operational cost
Scaling	Database-specific behaviors	Risk of misprediction

5.4 Practical Implications

This pipeline can be used by cloud teams with operators or orchestrators of Kubernetes or cloud-native.

Figure 3: Resource Scaling Decision Graph



6. Conclusion

Introduction of polyglot cloud database ecosystems have radically changed the way in which contemporary organizations design, scale and tune their data infrastructures. But this flexibility brings with it increasing problems in demand prediction, resource optimization, and cost control of the cloud under dynamically changing loads. This research suggested a holistic machine-learning-based framework that aligns workload forecasting, cost modeling, and optimization into a single line of

operation that is specific to heterogeneous database systems. The proposed system was tested under comparative ML tests, cost efficiency tests and using simulated workloads and all these showed that the proposed system had quantifiable benefits in comparison to the baseline scaling methods.

The general lesson is also direct and effective: predictive intelligence is becoming a necessity to cope with the complexity of polyglot cloud databases. The human intuition, way of provisioning, or reactive autoscaling no longer works at the scale and speed of the modern data systems. Combining ML forecasting with cost-sensitive optimization would help organizations to become much more financially efficient, alleviate performance pinpoints, and maintain service-level guarantees even in the face of bursty or unpredictable demands.

Critically, the results highlight the fact that machine learning is not just another supportive resource toward cloud management it is capable of being a decision engine which coordinates database provisioning, cost governance, and performance tuning in a data-driven and coordinated way. Within the wider framework of distributed computing, the study helps to advance the new paradigm of autonomous cloud operations, in which algorithmic decision-making will become a key factor of reliability and economic sustainability of the system.

6.1 Overview of the most important findings

The study provides a number of important conclusions that contribute to the development of the concept and effective operation of the polyglot cloud databases:

ML Models have a high predictive accuracy: Both Transformer-based and LSTM architectures obtained high fidelities in predicting trends of future resource utilization in heterogeneous database engines. They performed better than classical time-series models, especially when the workload is multi-modal and there is high-variance, such as polyglot environments.

Cost Optimization Delivers Significant Financial Benefits:ML predictions instead of reactive thresholds trained within the optimization module saved the average monthly costs by some 15 percent. What is more important is that it reduced cost variability which provided the organizations with more predictability in financial planning.

Capacity Usage Enhanced with no tradeoffs on SLAs: The system has decreased over-provisioning by almost 30 percent by proactively decreasing provisioning levels, but it did not decrease or harm performance metrics like latency and throughput. The level of SLA compliance increased to 92 percent (baseline) to 98 percent.

Predictive Scaling is an advantage to Polyglot Workload Patterns: Mixed read/write loads with document, relational and wide-column store, had the most improvement- confirming the argument that polyglot persistence increases scaling unpredictability.

Centralized Optimization Framework is Better than Isolated Scaling Models:A centralized approach to prediction and optimization of all database types, instead of treating each engine as an independent entity, gave more consistent results, as well as decreased scaling anomalies.

Realistic Yet Synthetic Workload Modeling is effective: The control of experimentation and the reflection of real workload dynamics of an enterprise were made possible by the use of synthetic yet structurally representative datasets.

In general, the findings confirm the viability and benefits of an ML-based capacity planning system that can serve to deal with the specific issues of heterogeneous cloud databases.

6.2 Future Research Directions.

Although it is a solid foundation, there are a number of potential avenues through which the framework can be further developed and its practical implications can be improved:

Implementation of Multi-Cloud and Hybrid-Clouds: Future studies must expand the logic of optimization to the environment that would include AWS, Azure, GCP, and on-premises data centers. Multi-cloud environments add new parameters such as inter-cloud latency, data transfer cost, redundancy policies, among them, which cannot be modeled without advanced modeling.

Reinforcement Learning of Autonomous and Continuous Optimization: The RL agents have the potential to substitute the fixed optimization policies and acquire an optimal policy of provisioning over time. This change would allow the self-tuning cloud infrastructures that can adapt to the changing workload conditions in real time.

Addition of Real Workloads of Production: Synthetic datasets have control and reproducibility, but real-world traces (e.g., e-commerce, IoT, telecom system) would also be a better assessment of the robustness of the system.

Pareto optimization model to cost-performance trade-off modeling: Organizations are known to have a clash of goals: reduce cost and maximize throughput and reduce latency. The evolutionary algorithms or multi-objective optimization of a Pareto frontier would provide a more refined decision-making.

Fine-Grained, Query-Based Prediction Models: Rather than simulation of resource utilization at the instance or cluster scale, future work can simulate resource needs at the query scale, which is especially useful to workloads that cross transaction and analytical engines.

Polystore Query Optimizers Interoperability: Predictive understanding may be important in systems such as Polystore++ and Polypheny-DB in directing queries or execution paths of queries between engines.

Security-Sensitive Scaling and Cost Modeling: Since increased use of encryption, zero-trust control, and data residency constraints are becoming part of organizations, future models need to capture the cost and performance implications of security settings.

Elucidating Models and Making Decisions Behaviors Transparent: Infrastructure decisions made by MLs bring about the governance issue. By incorporating explainable AI (XAI) methods, operators would be in a position to know the reason why a given model suggests some provisioning actions.

Through the discussion of these avenues, further studies can bring the field closer to complete autonomy, security, and cost-optimal polyglot cloud infrastructures.

References

1. Nikita, K. (2020). Managing and Optimising IoT Data and ML applications dependencies (Doctoral dissertation, Aalto University).
2. Gessert, F., Wingerath, W., & Ritter, N. (2020). Polyglot persistence in data management. In Fast and Scalable Cloud Data Management (pp. 149-174). Cham: Springer International Publishing.
3. Glake, D., Kiehn, F., Schmidt, M., Panse, F., & Ritter, N. (2022). Towards Polyglot Data Stores--Overview and Open Research Questions. arXiv preprint arXiv:2204.05779.
4. Joshi, S. (2022). Designing a Scalable Architecture for Ensemble Machine Learning & Collaboration.
5. Slott, H. (2017). 11 th Cloud Control Workshop.
6. Lau, M., Trivedi, S., He, Z., Pham, T., Perez, L., & Chakravorty, D. (2021). Research Cloud Bazaar: A software defined cloud workflow cost management tool. In Practice and Experience in Advanced Research Computing 2021: Evolution Across All Dimensions (pp. 1-4).
7. Eapen, B. R., Sartipi, K., & Archer, N. (2020). Serverless on FHIR: Deploying machine learning models for healthcare on the cloud. arXiv preprint arXiv:2006.04748.
8. Nagarathna, N. (2020, September). Remote Monitoring Solution with Predictive Analysis for Health Care Devices. In 2020 International Conference on Smart Electronics and Communication (ICOSEC) (pp. 682-689). IEEE.
9. Kodakandla, N. (2021). Serverless architectures: A comparative study of performance, scalability, and cost in cloud-native applications. Iconic Research and Engineering Journals, 5(2), 136-150.
10. Kusuma, P. (2022). A Holistic Framework for Designing Secure, Scalable, and Cost-Effective Cloud-Based E-Commerce Platforms. Journal of Advances in Cybersecurity Science, Threat Intelligence, and Countermeasures, 6(12), 7-16.
11. Elger, P., & Shanaghy, E. (2020). AI as a Service: Serverless machine learning with AWS. Manning.
12. Gollapudi, S. (2016). Practical machine learning (pp. 4-14). Birmingham: Packt Publishing.
13. Robertson, J., Fossaceca, J. M., & Bennett, K. W. (2021). A cloud-based computing framework for artificial intelligence innovation in support of multidomain operations. IEEE Transactions on Engineering Management, 69(6), 3913-3922.
14. Duggineni, S. (2022). The Synergy between Business Process and Big Data. Journal of Artificial Intelligence & Cloud Computing, 1(4), 1-7.
15. Soomro, K., Bhutta, M. N. M., Khan, Z., & Tahir, M. A. (2019). Smart city big data analytics: An advanced review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(5), e1319.
16. Zahid, H., Mahmood, T., Morshed, A., & Sellis, T. (2019). Big data analytics in telecommunications: literature review and architecture recommendations. IEEE/CAA Journal of Automatica Sinica, 7(1), 18-38.
17. Zhong, Z., & Buyya, R. (2020). A cost-efficient container orchestration strategy in kubernetes-based cloud computing infrastructures with heterogeneous resources. ACM Transactions on Internet Technology (TOIT), 20(2), 1-24.

18. Duvvuri, V. (2020). Minerva: A portable machine learning microservice framework for traditional enterprise SaaS applications. arXiv preprint arXiv:2005.00866.
19. Sharma, V. (2018). The cloud-based demand-driven supply chain. John Wiley & Sons.
20. Neeli, S. S. S. (2019). The Significance of NoSQL Databases: Strategic Business Approaches and Management Techniques. J. Adv. Dev. Res, 10(1), 11.
21. Kumar, V. S., Cuddihy, P., & Aggour, K. S. (2019, June). NodeGroup: a knowledge-driven data management abstraction for industrial machine learning. In Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning (pp. 1-4).
22. Eppinger, F., & Störl, U. (2022). Nosql database tuning through machine learning. arXiv preprint arXiv:2212.12301.
23. Mazumdar, S., Seybold, D., Kritikos, K., & Verginadis, Y. (2019). A survey on data storage and placement methodologies for cloud-big data ecosystem. Journal of Big Data, 6(1), 1-37.
24. Sethupathy, A., & Kumar, U. (2020). Cloud-Native Architectures for Real-Time Retail Inventory and Analytics Platforms. International Journal of Novel Research and Development, 5, 339-355.
25. Vogt, M., Stiemer, A., & Schuldt, H. (2018, December). Polypheny-DB: towards a distributed and self-adaptive polystore. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 3364-3373). IEEE.
26. Traukina, A., Thomas, J., Tyagi, P., & Reddipalli, K. (2018). Industrial Internet Application Development: Simplify IIoT Development Using the Elasticity of Public Cloud and Native Cloud Services. Packt Publishing Ltd.
27. Drabas, T., & Lee, D. (2017). Learning PySpark. Packt Publishing Ltd.
28. Misargopoulos, A., Papavassiliou, G., Gizelis, C. A., & Nikolopoulos-Gkamatsis, F. (2021, June). TYPHON: hybrid data lakes for real-time big data analytics–an evaluation framework in the telecom industry. In IFIP International Conference on Artificial Intelligence Applications and Innovations (pp. 128-137). Cham: Springer International Publishing.
29. Singhal, R., Zhang, N., Nardi, L., Shahbaz, M., & Olukotun, K. (2019, July). Polystore++: Accelerated polystore system for heterogeneous workloads. In 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS) (pp. 1641-1651). IEEE.
30. Nikolopoulos-Gkamatsis, F. (2021, June). TYPHON: Hybrid Data Lakes for Real-Time Big Data–An Evaluation Framework Analytics in the Telecom Industry. In Artificial Intelligence Applications and Innovations. AIAI 2021 IFIP WG 12.5 International Workshops: 5G-PINE 2021, AI-BIO 2021, DAAI 2021, DARE 2021, EEAI 2021, and MHDW 2021, Hersonissos, Crete, Greece, June 25–27, 2021, Proceedings (Vol. 628, p. 128). Springer Nature.