

Resource Management in AI-Enabled Cloud Native Databases: A Systematic Literature Review Study

Shantanu Kumar, Shruti Singh, Harshavardhan Nerella

Submitted: 06/02/2024 Revised: 14/03/2024 Accepted: 22/03/2024

Abstract: In the face of digital transformation, organizations are currently grappling with the differences between new and classic business models, which require agile, adaptable, personalized, and intelligent approaches. Cloud-native, as a novel technological idea, plays a crucial role in assisting organizations in constructing and operating adaptable and expandable applications in modern dynamic environments, including public, private, and hybrid clouds. Cloud-native offers several benefits, including enhanced performance, optimized resource utilization, reduced operational expenses, and improved scalability. The main aim of this work is to investigate resource management in cloud-native databases that are enabled by artificial intelligence. The research methodology utilized in this study is a comprehensive literature review. This study conducted a comprehensive analysis of 30 scholarly papers sourced from various online academic databases spanning the period from 2018 to 2024. Based on this research, efficiently managing resources in AI-enabled cloud-native databases is crucial for enterprises seeking to harness the potential of artificial intelligence while optimizing productivity and reducing expenses. The results of this study assist database developers in choosing suitable objectives and devising strategies for enhancing AI-enabled Cloud-native databases.

Keywords: Resource Management; Artificial Intelligence; Cloud Native Databases; Digital Transformation.

Introduction:

Cloud-native databases (CNDBs) have gained significance in cloud computing as they fulfil the requirements of elasticity, scalability, management, and on-demand usage for various applications [1]. These issues posed by cloud applications offer new prospects for CNDBs that conventional on premise enterprise database systems are unable to adequately tackle. Common Network Database possess several characteristics, including the ability to support several tenants, the separation of computing and storage resources, and the use of logs as the underlying database. These characteristics enhance the flexibility and expandability of the database, thereby resolving multiple difficulties outlined before [2]. Like traditional databases, CNDBs also utilize Artificial Intelligence (AI) technologies to

optimize database performance. The application of AI technology in the characteristics of CNDBs significantly enhances the database's performance. Amidst the ongoing advancements in AI technology, AI-empowered CNDBs are becoming increasingly popular. Currently, the progress of CNDBs is at an early stage, and the utilization of AI technology in CNDBs is not yet fully developed [3]. In the absence of a maturity model for AI-empowered CNDBs, database developers may experience uncertainty over the implementation of AI technology, while database users may face confusion when choosing CNDBs. Consequently, there is a need for a maturity model to evaluate AI-empowered CNDBs. This model will serve as a tool to analyse the current state of CNDBs and provide guidance for incremental enhancements [2].

1Senior Software Engineer, Amazon, US

Email ID: reach.shantanuk@gmail.com

2Program Manager, Washington State University, US

Email ID: sshruti.connect@gmail.com

3Senior Cloud Engineer, MassMutual, US

Email ID: nerellaharshavardhan@outlook.com

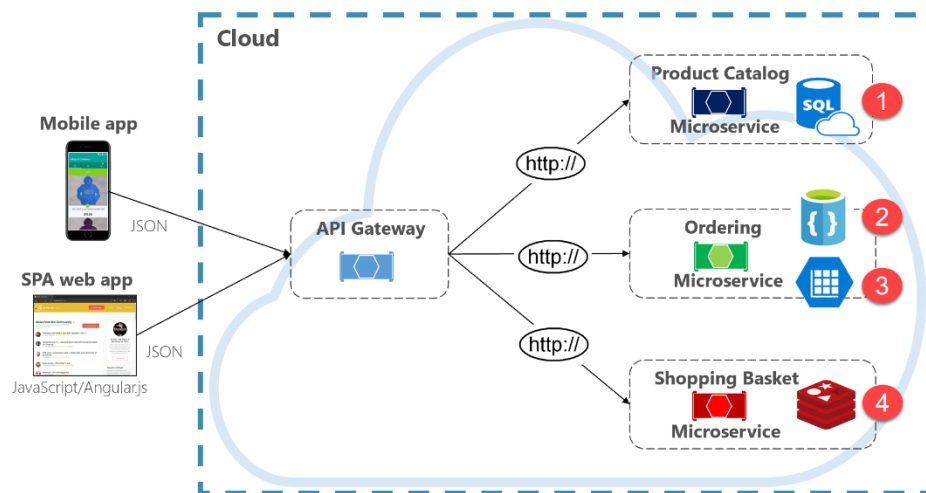


Fig 1: Basic Cloud Native Databases Pattern [4]

Scope of the study:

This study performs a comprehensive assessment of existing literature to investigate the complexities of managing resources in cloud-native databases that utilize artificial intelligence and are based on maturity models. This study examines the methods, approaches, and instruments used to efficiently allocate and optimize resources in these ever-changing situations. The study aims to deliver a comprehensive overview of existing research, identify key trends, challenges, and best practices, and offer insights for future advancements in the critical area of database management by examining the intersection of maturity models, artificial intelligence, and cloud-native database technologies.

Objectives of this study:

- To systematically review the literature to identify and analyse existing resource management approaches within maturity model-based AI-enabled cloud-native databases.
- To evaluate the effectiveness and limitations of the identified resource management approaches.
- To identify opportunities for advancement and future research directions in resource management within maturity model-based AI-enabled cloud-native databases.

Research questions:

- What resource management strategies are in place for cloud-native databases with AI enabled that are based on maturity models?
- What are the positive aspects and drawbacks of the different resource management strategies?

- What are the distinct possibilities for development and future research directions in resource management in cloud-native databases with AI enabled through maturity models?

Research Background:

Efficiently managing resources in AI-enabled cloud-native databases based on maturity models is a crucial aspect of optimizing databases in modern IT environments [5]. Cloud-native databases require advanced resource allocation algorithms to provide optimal performance and cost-effectiveness by using the scalability and flexibility of cloud infrastructures. Simultaneously, maturity models provide as standards for assessing the developmental advancement of database management techniques, directing enterprises towards more effective and robust systems. Artificial intelligence enhances these efforts by introducing predictive analytics, autonomous decision-making, and adaptive resource allocation capabilities [6]. Nevertheless, this merging of technologies also presents complex obstacles, such as the necessity to traverse the ever-changing nature of cloud settings, strike a balance between performance requirements and resource expenses, and utilize AI capabilities wisely to improve database efficiency [7]. Gaining a deep understanding of the intricacies of resource management in this particular situation is crucial for fostering innovation, enhancing database efficiency, and fulfilling the changing requirements of contemporary data-centric businesses [7], [8]. This systematic literature study is necessary to fully comprehend the present condition of resource management techniques, pinpoint any deficiencies and constraints, and lay the groundwork for future

research and innovation in this crucial field of database administration.

Methodology:

The study employs the Systematic Literature Review methodology. This study utilizes the PRISMA methodology to investigate resource

management in AI-enabled cloud-native databases. PRISMA, or Preferred Reporting Items for Systematic Reviews and Meta-Analyses, provides a systematic framework for identifying relevant research, extracting important data, and synthesizing findings to inform evidence-based decision-making [9].

Table 1: Relevant publications were identified from internet repositories in this study.

DIGITAL LIBRARY	URL
Scopus	https://www.scopus.com/sources.uri?zone=TopNavBar&origin=searchbasic
Semantic Scholar	https://www.semanticscholar.org/
IEEE Xplore	https://ieeexplore.ieee.org/Xplore/home.jsp
Science Direct	https://www.sciencedirect.com/
Springer	https://link.springer.com/
Web of Science	https://wosjournal.com/
PubMed	https://pubmed.ncbi.nlm.nih.gov/

The search strings are formulated to encompass a wide spectrum of pertinent literature pertaining to resource management in maturity model-based AI-enabled cloud-native databases. This includes research on resource allocation strategies, maturity models, integration of artificial intelligence, optimization techniques, and performance tuning in cloud-native database environments. Modify and perfect these search queries as necessary according to your own study goals and interests. The search strings are *Cloud Native Databases* OR Cloud-Native Database Management * AND Resource Management* OR Resource Allocation *AND*

Maturity Model OR Maturity Framework AND* Artificial Intelligence* OR AI * AND Optimization.* The strings were used to extract the title and abstract sections. Only papers released between 2018 and 2024 were selected for assessment to concentrate on the most recent advancements in the area.

Inclusion/Exclusion Criteria:

The following table presents the potential criteria for inclusion and exclusion in this systematic literature review on the investigation of performance and scalability in cloud object stores:

Table 2: Inclusion/Exclusion Criteria

Criteria	Inclusion	Exclusion
Type of Study	Academic Research papers and Review articles	Editorials and opinions
Publication date	Studies published within the last 5 years	Studies published over 5 years ago
Language	English	Non-English studies
Subject Are	Cloud computing, maturity model-based AI-enabled cloud-native databases, resource management.	Irrelevant Subjects
Access Availability	Open access studies	Studies behind paywalls or lacking access

Peer – Review	Peer-reviewed studies	Non-peer-reviewed studies
---------------	-----------------------	---------------------------

The information was gathered by creating and assessing the search term across multiple well-known databases. The references were filtered to remove any instances of duplication once they were imported using Endnote. Consequently, there were 110 papers left in all, each completely unique from the others. After a comprehensive analysis of abstracts and titles, this study was able to locate papers that had the chosen keywords. The generated references were then transferred to an Excel spreadsheet so that additional filtering and analysis could be done. The authors, the year of publication, the title, and the abstract were all listed in this spreadsheet. There are a total of 79 research

publications in the Excel spreadsheet. Excel was used to review all of the paper abstracts, with an emphasis on those that closely related to the goals of the research. 51 articles in total were initially located; however, some of them were removed because they were categorized as books, books, or portions of books, or as grey literature. Furthermore, some of the articles were taken out of circulation since they could not be downloaded. The final shortlist consisted of thirty papers. The method used to choose articles for the SLR during a literature search is depicted in a flow diagram (Figure below) that is grounded in the PRISMA Flow Diagram.

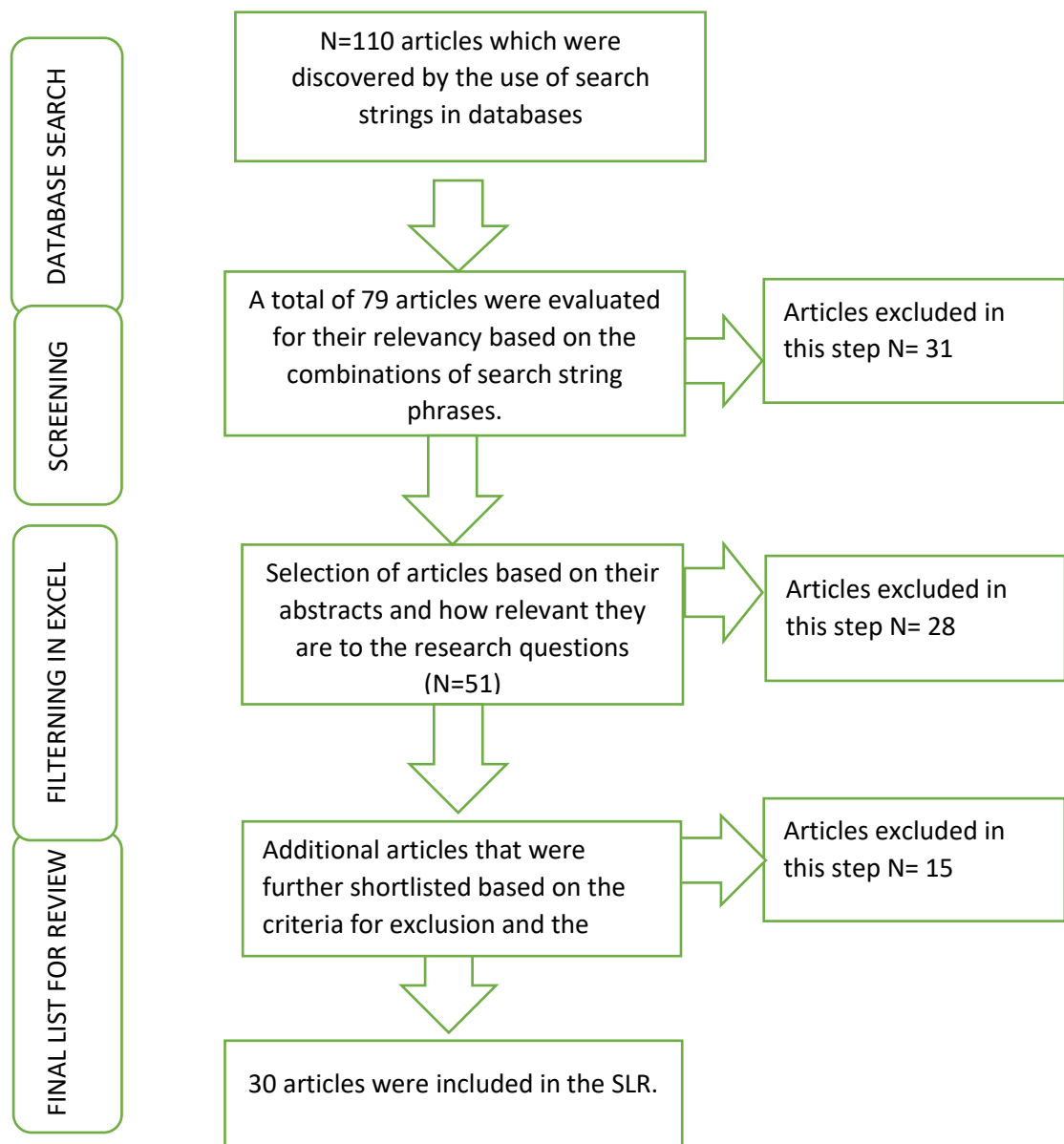


Fig 3: Literature search for SLR publications (based on PRISMA flow diagram).

Results And Discussions:

CNDBs can be classified into two main branches. Some databases, like CockroachDB, TiDB, and YugabyteDB, are built on the foundation of Spanner. The other databases are built on the Aurora platform, including Socrates, PolarDB, CynosDB, ArkDB, TarusDB, and others. These databases possess distinct characteristics, yet the majority of their features are identical [2]. Cloud-native apps (CNA) are software programs that inherit advantageous cloud computing attributes. They utilize cloud platform alongside infrastructure services to make themselves operational, provide their own functionality in the shape of software service interfaces, are resistant to dependency service unavailability alongside other incidents, scale flexibly with user requests, are consistently accessible upon request, and are charged based on a pay-per-use utility scheme without any initial costs [10]. The natural focus on providing services in CNA necessitates the use of a micro services

architecture, which includes both stateful and stateless services. Data handling is limited to the stateful services. These systems must also possess a high level of availability and resilience to effectively prevent any loss, corruption, or delay in data processing. Prior studies have examined databases, alongside message queues, alongside key-value stores, file systems, and other data access architectures in relation to these needs [11]. The desired attributes rely on the ability to quickly replicate services [12], which necessitates the use of consistent data replication and sharing methods.

The table below provides a comparison of various CNDBs, highlighting their shared characteristics such as multi-tenancy, disaggregation of compute and storage, cross availability zone/region support, near-data processing, utilization of logs as the database, and employment of distributed alongside shared memory.

Table 3: Cloud-native databases

AUTHORS AND YEAR	DATABASES	EXPLANATION
Bacon et al., (2017) [13]	Spanner	Spanner is a highly scalable and globally distributed database that was specifically developed, constructed, and implemented by Google. The design of this system allows it to easily accommodate millions of machines distributed over hundreds of datacentres, as well as handle trillions of database rows.
Cao et al., (2021) [14]	TiDB	TiDB is a database that uses the Raft consensus algorithm and supports Hybrid Transactional/Analytical Processing (HTAP). The system features a multi-Raft storage system comprising a row store and a column store. The system has the capability to construct a SQL engine that can handle extensive distributed transactions and resource-intensive analytical queries.
Verbitski et al., (2017) [15]	Amazon Aurora	Amazon Aurora is a database service designed specifically for handling online transaction processing (OLTP) workloads. Aurora delegates the task of redo processing to a multi-tenant scale-out storage service specifically designed for Aurora. This not only minimizes network traffic, but also enables quick crash recovery, failovers to replicas without data loss, and ensures fault-tolerant and self-healing storage.
Cao et al., (2020) [16]	POLARDB	POLARDB is a recently developed cloud-native OLTP database specifically created by Alibaba Cloud. The database computing nodes and storage nodes are interconnected via a high-speed Remote Direct Memory Access (RDMA) network. In order to guarantee a high level of accessibility, POLARDB employs the Parallel-Raft protocol to replicate data among the storage nodes by creating three copies.
Depoutovitch et al., (2020) [17]	TaurusDB	POLARDB is a recently developed cloud-native OLTP database specifically created by Alibaba Cloud. The database computing nodes and storage nodes are interconnected via a high-speed Remote Direct Memory Access network. In order to guarantee a high level of accessibility, POLARDB employs the Parallel-Raft protocol to replicate data among the storage nodes by creating three copies.

Resource management in AI -enabled cloud native databases:

Resource management in AI-enabled cloud-native databases focuses on improving the allocation and usage of computing resources, including CPU, memory, storage, and network bandwidth, to efficiently execute AI tasks [18]. This encompasses dynamic scaling techniques to adapt to fluctuating demands, sophisticated load balancing to evenly distribute jobs among nodes, and predictive analytics to anticipate resource requirements and proactively change provisioning. In cloud environments, the use of advanced techniques like as containerization and orchestration allows for easy deployment and control of database instances. This

improves scalability, dependability, and cost-effectiveness. Organizations may optimize speed and scalability while reducing operational costs in maintaining their cloud-native database infrastructure by utilizing AI algorithms and automation [19].


Moreover, the performance of CNDBs, which are specifically developed for cloud architecture, is primarily determined by the efficiency of cloud resource management. The categorization of resource management includes four main categories: resource prediction, alongside resource scheduling, alongside resource control, alongside resource scaling. The details are provided in the table below.

Table 4: AI-enabled resource management


AUTHORS AND YEARS	TYPE OF RESOURCE MANAGEMENT	DESCRIPTION
Wang et al., (2021) [20]	Resource Prediction	The system, called TPC (Trend Prediction based on Clustering), offers a fast method for predicting KPI trends. Its purpose is to assist the operation and maintenance team in making timely and appropriate adjustments to cloud resources.
Tan et al., (2019) [21]	Resource Scheduling	iBTune is designed to autonomously coordinate the optimization of buffer pool tuning across all database instances.
Zhang et al., (2021)	Resource Control	ResTune is designed to improve resource consumption while ensuring that SLA limitations regarding throughput and latency requirements are not violated.
Salmanian et al., (2022)	Resource Scaling	The proposal introduces a Hybrid Auto-Scaler (HAS) that automatically adjusts the necessary resources to meet the demand of the application. HAS allocates the expected resources by calculating the necessary capacity. Additionally, it adjusts the allocation of resources when the currently given resources are insufficient to meet the present demands.

Case studies as examples:

The following is an explanation of a few of case studies that demonstrate how organizations have implemented resource management in cloud-native databases that are empowered with artificial intelligence.

 *Netflix:* Netflix exemplifies a firm that extensively depends on AI-powered cloud-native databases for functions such as content recommendation, personalization, and other machine learning activities. Netflix encounters substantial obstacles in properly managing resources to provide personalized recommendations in real-time and optimize expenses, given its extensive user base and enormous content catalogue [18]. Netflix

required the ability to adjust its infrastructure in real-time to accommodate varying workloads and provide customized suggestions to millions of users throughout the globe. This organization used a cloud-native database system that incorporates AI-powered resource management features. Netflix utilizes sophisticated algorithms for workload prediction, auto-scaling, and intelligent resource allocation to optimize resource use according to demand, guaranteeing optimal performance and cost-efficiency. Netflix utilizes AI-enabled resource management to provide consumers with personalized recommendations in real-time, while also optimizing infrastructure costs. The organization attains exceptional scalability and dependability, guaranteeing a smooth streaming experience for its consumers.

 *Uber:* Uber utilizes AI-enabled cloud-native databases to operate its ride-sharing network, managing millions of transactions and requests on a daily basis. Efficient resource management is necessary for Uber's platform to handle fluctuating workloads, guarantee low-latency answers, and optimize costs. Uber required the ability to expand its database infrastructure in a flexible manner to accommodate high levels of traffic during busy periods and special occasions, while also keeping expenses to a minimum [24]. This group successfully deployed a database solution that is specifically designed for cloud environments and incorporates advanced artificial intelligence algorithms to efficiently manage system resources. Uber utilizes machine learning algorithms to estimate workloads, automatically adjust resource capacity, and intelligently route services. This optimization strategy ensures that resources are efficiently utilized according on demand, resulting in both high availability and cost efficiency. By utilizing AI-powered resource management, Uber is able to effectively manage high levels of traffic, guaranteeing quick response times for ride requests while lowering expenses related to infrastructure. The organization attains scalability and reliability, ensuring a smooth and uninterrupted experience for both riders and drivers.

Resource allocation and utilization in cloud-native databases enhanced with artificial intelligence technology pose various obstacles and offer potential advantages in terms of efficiency and effectiveness. They are:

- Effectively managing resources in a dynamic and distributed environment with shifting demands can provide significant complexity. To achieve the most efficient distribution and usage of resources, it is necessary to employ advanced algorithms and techniques that take into account aspects such as data localization, network latency, and workload diversity [25].
- Cloud-native databases are specifically engineered to dynamically adjust their capacity in response to changing demand. Nevertheless, efficiently overseeing resources over an expanding number of nodes and clusters while upholding performance and dependability poses a substantial difficulty. Scaling mechanisms need to be both efficient and capable of quickly adapting to sudden increases or decreases in workload [26].
- Although cloud resources have the advantage of scalability, they also come with the disadvantage of incurring expenditures. Efficient resource allocation entails achieving a balance between performance demands and cost factors. Efficiently managing resource allocation to reduce costs while still fulfilling performance service level agreements (SLAs) is crucial for maintaining cost-effective operations [26].
- AI tasks frequently need substantial computational resources. It is vital to ensure that resources are allocated and used efficiently in order to meet performance criteria. This involves enhancing the execution of queries, processing of data, and inference of models in order to reduce latency and optimize throughput.
- Cloud-native databases efficiently manage substantial amounts of data that are spread across several nodes. Efficient resource management entails guaranteeing the availability, uniformity, and resilience of data while improving the arrangement and retrieval methods to reduce delay and maximize data transfer rate.
- It is not feasible to manage resources manually on a large scale. Automation and orchestration technologies are crucial for the dynamic provisioning, scaling, monitoring, and optimization of resource utilization. Utilizing AI algorithms for predictive analytics and autonomous decision-making can significantly improve the efficiency of resource management [25].

Optimization of AI-enabled cloud native database resource management in organizations:

It is essential for businesses to optimize resource management in cloud-native databases that are empowered with artificial intelligence in order to achieve efficient operations, reduce expenses, and deliver optimal performance from their databases. The following are many ways that can be utilized to optimize resource management in environments like these [27]:

Predictive Analytics: Utilize predictive analytics methodologies to anticipate resource utilization and workload trends. Through the examination of past data and patterns, businesses can predict future resource requirements and strategically distribute resources to meet projected workloads, thus preventing insufficient or excessive provisioning.

Auto Scaling: Employ auto-scaling capabilities to adaptively modify resource allocation in response to immediate changes in workload intensity. Develop algorithms that autonomously adjust the quantity of database instances, computational resources, or storage capacity based on fluctuations in demand, guaranteeing optimal performance and cost-effectiveness.

Intelligent workload distribution: Deploy sophisticated workload distribution algorithms to evenly allocate duties among database nodes and clusters. Employ load-balancing algorithms to direct queries and requests to the most suitable nodes, taking into account parameters such as node capacity, network latency, and data locality. This approach optimizes the utilization of resources and reduces response times.

Cost optimization: Maximize resource efficiency to decrease expenses while maintaining performance and dependability. Develop and apply scheduling policies that prioritize workloads based on their cost-to-benefit ratio, ensuring that expensive resources are only given to mission-critical jobs while less important workloads are assigned lower-cost resources [28].

Performance Tuning: Regularly observe and adjust database settings, query execution strategies, and data retrieval methods to enhance performance. Discover and resolve performance limitations by utilizing methods such as optimizing queries, tweaking indexes, and redesigning database schemas. This will enhance data processing speed and reduce delays to a minimum.

AI driven resource allocation: Incorporate artificial intelligence and machine learning algorithms into resource management procedures to facilitate independent decision-making and optimization. Utilize AI-driven methodologies, like reinforcement learning and genetic algorithms, to consistently acquire knowledge from previous encounters and adjust resource allocation tactics to changing workload patterns, resulting in the attainment of self-optimizing database systems [28].

Intelligent Load balancing: Deploy sophisticated load-balancing techniques to evenly spread workloads across database nodes and clusters. These algorithms take into account parameters such as the capacity of nodes, the latency of the network, alongside the localization of data to optimize the use of resources along with reduce reaction times [29].

Policy – Based Resource Allocation: Develop resource allocation systems that enable businesses to establish rules and regulations for allocating resources based on criteria such as workload priorities, service-level agreements (SLAs), and cost limitations. These policies guarantee that resources are distributed in line with the goals and needs of the organization [30].

Conclusion:

Resource management in AI-enabled cloud-native databases is a complex task that is essential for maximizing performance, scalability, and cost efficiency. By utilizing auto-scaling mechanisms, predictive analytics, intelligent load balancing, cost optimization algorithms, containerization, orchestration, AI-driven resource allocation, and policy-based strategies, organizations can effectively adjust to changing workloads and make efficient use of computing resources. These technologies enable enterprises to fully utilize artificial intelligence in cloud-native database settings, assuring smooth operations, improved scalability, and efficient resource allocation. In order to remain competitive and take advantage of the revolutionary possibilities of cloud-native databases, enterprises must implement robust resource management strategies as artificial intelligence (AI) continues to drive innovation and redefine sectors.

References:

- [1].Li, F. (2019). Cloud-native database systems at Alibaba: Opportunities and challenges. *Proceedings of the VLDB Endowment*, 12(12), 2263-2272.
- [2].Feng, X., Guo, C., Jiao, T., & Song, J. (2022). A maturity model for AI-empowered cloud-native databases: from the perspective of resource management. *Journal of Cloud Computing*, 11(1), 39.
- [3].Ton That, D. H., Wagner, J., Rasin, A., & Malik, T. (2019). PLI⁺⁺: efficient clustering of cloud databases. *Distributed and Parallel Databases*, 37, 177-208.
- [4].Cloud-native data patterns. (2022). Retrieved April 27, 2024, from Microsoft.com website: <https://learn.microsoft.com/en-us/dotnet/architecture/cloud-native/distributed-data>.
- [5].Zeb, S., Rathore, M. A., Mahmood, A., Hassan, S. A., Kim, J., & Gidlund, M. (2021, December). Edge intelligence in softwarized 6G: Deep learning-enabled network traffic predictions. In *2021 IEEE Globecom Workshops (GC Wkshps)* (pp. 1-6). IEEE.
- [6].Zhang, R., Li, Y., Li, H., & Wang, Q. (2022). Evolutionary game analysis on cloud providers and enterprises' strategies for migrating to cloud-native under digital transformation. *Electronics*, 11(10), 1584.
- [7].Samdanis, K., & Taleb, T. (2020). The road beyond 5G: A vision and insight of the key technologies. *IEEE Network*, 34(2), 135-141.
- [8].Cardoso, K. V., Both, C. B., Prade, L. R., Macedo, C. J., & Lopes, V. H. L. (2020). A softwarized perspective of the 5G networks. *arXiv preprint arXiv:2006.10409*.
- [9].Mengist, W., Soromessa, T., & Legese, G. (2020). Method for conducting systematic literature review and meta-analysis for environmental science research. *MethodsX*, 7, 100777.
- [10].Spillner, J., Toffetti, G., & López, M. R. (2018). Cloud-Native Databases: An Application Perspective. In *Advances in Service-Oriented and Cloud Computing: Workshops of ESOC 2017, Oslo, Norway, September 27-29, 2017, Revised Selected Papers 6* (pp. 102-116). Springer International Publishing.
- [11].Szczyrbowski, M., & Myszor, D. (2015, May). Comparison of the behaviour of local databases and databases located in the cloud. In *International Conference: Beyond Databases, Architectures and Structures* (pp. 253-261). Cham: Springer International Publishing.
- [12].Li, G., Dong, H., & Zhang, C. (2022). Cloud databases: New techniques, challenges, and opportunities. *Proceedings of the VLDB Endowment*, 15(12), 3758-3761.
- [13].Bacon, D. F., Bales, N., Bruno, N., Cooper, B. F., Dickinson, A., Fikes, A., ... & Woodford, D. (2017, May). Spanner: Becoming a SQL system. In *Proceedings of the 2017 ACM International Conference on Management of Data* (pp. 331-343).
- [14].Cao, Y., Dong, Q., Wang, D., Liu, Y., Zhang, P., Yu, X., & Niu, C. (2021). TIDB: a comprehensive database of trained immunity. *Database*, 2021, baab041.
- [15].Verbitski, A., Gupta, A., Saha, D., Brahmadesam, M., Gupta, K., Mittal, R., ... & Bao, X. (2017, May). Amazon aurora: Design considerations for high throughput cloud-native relational databases. In *Proceedings of the 2017 ACM International Conference on Management of Data* (pp. 1041-1052).
- [16].Cao, W., Liu, Y., Cheng, Z., Zheng, N., Li, W., Wu, W., ... & Zhang, T. (2020). {POLARDB} meets computational storage: Efficiently support analytical workloads in {Cloud-Native} relational database. In *18th USENIX conference on file and storage technologies (FAST 20)* (pp. 29-41).
- [17].Depoutovitch, A., Chen, C., Chen, J., Larson, P., Lin, S., Ng, J., ... & He, Y. (2020, June). Taurus database: How to be fast, available, and frugal in the cloud. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (pp. 1463-1478).
- [18].Toffetti, G., Brunner, S., Blöchliger, M., Spillner, J., & Bohnert, T. M. (2017). Self-managing cloud-native applications: Design, implementation, and experience. *Future Generation Computer Systems*, 72, 165-179.
- [19].Li, F. (2023). Modernization of databases in the cloud era: Building databases that run like Legos. *Proceedings of the VLDB Endowment*, 16(12), 4140-4151.
- [20].Wang, X., Li, N., Zhang, L., Zhang, X., & Zhao, Q. (2021, May). Rapid Trend Prediction for Large-Scale Cloud Database KPIs by Clustering. In *2021 IEEE/ACM International Workshop on Cloud Intelligence (CloudIntelligence)* (pp. 1-6). IEEE.
- [21].Tan, J., Zhang, T., Li, F., Chen, J., Zheng, Q., Zhang, P., ... & Zhang, R. (2019). ibtune: Individualized buffer tuning for large-scale cloud databases. *Proceedings of the VLDB Endowment*, 12(10), 1221-1234.

- [22].Zhang, X., Wu, H., Chang, Z., Jin, S., Tan, J., Li, F., ... & Cui, B. (2021, June). Restune: Resource oriented tuning boosted by meta-learning for cloud databases. In *Proceedings of the 2021 international conference on management of data* (pp. 2102-2114).
- [23].Salmanian, Z., Izadkhah, H., & Isazadeh, A. (2022). Auto-scale resource provisioning in IaaS clouds. *The Computer Journal*, 65(2), 297-309.
- [24].Velayutham, S., & Shanmugam, G. (2021). Artificial Intelligence assisted Canary Testing of Cloud Native RAN in a mobile telecom system.
- [25].Abouelyazid, M., & Xiang, C. (2019). Architectures for AI Integration in Next-Generation Cloud Infrastructure, Development, Security, and Management. *International Journal of Information and Cybersecurity*, 3(1), 1-19.
- [26].Zeb, S., Rathore, M. A., Hassan, S. A., Raza, S., Dev, K., & Fortino, G. (2023). Toward AI-enabled nextG networks with edge intelligence-assisted microservice orchestration. *IEEE Wireless Communications*, 30(3), 148-156.
- [27].Singh, A. (2023). Optimization of the Cloud-Native Infrastructure using Artificial Intelligence.
- [28].Gill, S. S., Xu, M., Ottaviani, C., Patros, P., Bahsoon, R., Shaghaghi, A., ... & Uhlig, S. (2022). AI for next generation computing: Emerging trends and future directions. *Internet of Things*, 19, 100514.
- [29].Schein, S., Arutiunian, G., Burshtein, V., Sadeh, G., Townshend, M., Friedman, B., & Sadr-azodi, S. (2021). Developing Medical AI: a cloud-native audio-visual data collection study. *arXiv preprint arXiv:2110.03660*.
- [30].Budhwar, P., Chowdhury, S., Wood, G., Aguinis, H., Bamber, G. J., Beltran, J. R., ... & Varma, A. (2023). Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. *Human Resource Management Journal*, 33(3), 606-659.