International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903016

Retrieval-Augmented Generation with Serverless Cloud Computing for Enterprise Knowledge Management and Decision Intelligence

Cynthiya Mohan

Project Manager, Amadeus Software labs, London Area, United Kingdom

Publication History: Received: 10.05.2026; Revised: 04.06.2026; Accepted: 10.06. 2026; Published: 23.06.2026.

ABSTRACT: Retrieval-Augmented Generation (RAG) combined with serverless cloud computing represents a transformative approach for improving enterprise knowledge management and decision intelligence by integrating large language models with dynamic organizational information repositories. Traditional enterprise knowledge systems often struggle with fragmented data sources, outdated information, and limited contextual understanding. RAG addresses these limitations by retrieving relevant information from enterprise databases, documents, and knowledge graphs before generating responses through artificial intelligence models. When deployed using serverless cloud architectures, RAG systems gain scalability, cost efficiency, and operational flexibility by dynamically allocating computational resources according to demand. This integration enables organizations to develop intelligent systems capable of providing accurate insights, supporting strategic decisions, and enhancing employee productivity. The research explores the role of serverless RAG frameworks in enterprise environments, focusing on architecture, implementation strategies, knowledge accessibility, and decision-support capabilities. The study adopts a conceptual research methodology based on analysis of existing literature, technological developments, and enterprise AI adoption practices. Findings indicate that serverless RAG solutions can improve organizational intelligence by reducing information retrieval barriers, enhancing contextual accuracy, and enabling real-time decision-making. However, challenges related to data security, model reliability, governance, and integration complexity remain important considerations. The study concludes that the combination of RAG and serverless computing provides a sustainable foundation for next-generation enterprise knowledge ecosystems.

KEYWORDS: Retrieval-Augmented Generation, Serverless Cloud Computing, Enterprise Knowledge Management, Decision Intelligence, Artificial Intelligence, Large Language Models, Cloud Architecture, Knowledge Retrieval, Organizational Analytics, Intelligent Decision Support

I. INTRODUCTION

The rapid growth of digital transformation has significantly increased the volume, diversity, and complexity of enterprise information. Organizations generate and store massive amounts of structured and unstructured data through business applications, customer interactions, operational systems, research documents, and communication platforms. Although this information represents a valuable strategic asset, many enterprises experience difficulties in effectively accessing and utilizing knowledge for decision-making. Conventional knowledge management systems frequently depend on keyword-based search mechanisms and manually maintained repositories, which often fail to provide contextual understanding and adaptive responses. The emergence of artificial intelligence, particularly large language models (LLMs), has created new opportunities for improving enterprise knowledge utilization. However, standalone LLMs have limitations, including dependence on training data, inability to access proprietary organizational information, and risks of generating inaccurate responses.

Retrieval-Augmented Generation provides a solution by combining information retrieval techniques with generative artificial intelligence. Instead of relying solely on pre-trained model knowledge, RAG systems retrieve relevant information from external knowledge sources and incorporate that information into the generation process. This approach enables AI systems to provide more accurate, context-aware, and organization-specific responses. In enterprise environments, RAG can connect employees with internal documents, policies, reports, customer records, and operational knowledge, creating intelligent assistants that support research, analysis, and decision-making.

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903016

Serverless cloud computing further enhances the capabilities of RAG systems by providing an elastic and efficient deployment model. Unlike traditional cloud infrastructure requiring organizations to manage servers, operating systems, and scaling mechanisms, serverless platforms automatically allocate resources based on workload requirements. This capability is particularly valuable for AI applications, where computational demands may fluctuate significantly. Enterprises can deploy retrieval pipelines, embedding services, vector databases, and AI inference components through serverless architectures, reducing infrastructure management responsibilities and optimizing operational costs.

The integration of RAG with serverless computing has important implications for enterprise knowledge management and decision intelligence. Decision intelligence focuses on combining data, analytics, artificial intelligence, and business knowledge to improve organizational choices. By providing decision-makers with timely and relevant insights, RAG-powered systems can enhance strategic planning, risk assessment, customer analysis, and operational optimization. These systems can transform passive information repositories into active intelligence platforms capable of assisting users in complex problem-solving activities.

Despite its advantages, implementing RAG within enterprises presents several challenges. Organizations must address issues related to data privacy, access control, knowledge quality, model transparency, and compliance requirements. Enterprise data often contains sensitive information, requiring robust security mechanisms and governance frameworks. Additionally, retrieval accuracy directly affects generated responses, making effective document indexing, metadata management, and evaluation methods essential. Therefore, understanding the technological foundations and implementation strategies of serverless RAG systems is important for organizations seeking to achieve reliable AI-driven knowledge management.

II. LITERATURE REVIEW

This research examines how Retrieval-Augmented Generation supported by serverless cloud computing can enhance enterprise knowledge management and decision intelligence. The study analyzes existing research, identifies technological opportunities and limitations, and proposes methodological considerations for developing effective enterprise AI solutions.

The development of artificial intelligence-based knowledge management systems has attracted significant attention from researchers and industry practitioners. Earlier knowledge management approaches focused primarily on information storage, classification, and retrieval through databases and enterprise content management platforms. These systems improved organizational access to information but often lacked the ability to understand natural language queries and provide intelligent recommendations. Advances in machine learning and natural language processing introduced more sophisticated methods for extracting meaning from large-scale information resources.

Large language models have become a major advancement in natural language processing due to their ability to generate human-like text and perform complex reasoning tasks. Models based on transformer architectures demonstrated significant improvements in language understanding and generation. However, researchers identified limitations associated with knowledge freshness, domain specialization, and factual reliability. Because LLMs generate responses based on learned patterns rather than direct access to current organizational information, they may produce incorrect or unsupported outputs. These limitations have encouraged research into methods that combine generative models with external knowledge retrieval.

Retrieval-Augmented Generation was introduced as an approach to improve the reliability and relevance of language models by incorporating retrieved documents during response generation. Research has shown that RAG systems can reduce hallucination risks and improve domain-specific performance. In enterprise contexts, retrieval mechanisms allow AI applications to access internal knowledge sources without requiring complete retraining of language models. This capability reduces implementation complexity and enables organizations to customize AI systems according to specific business requirements.

The development of vector databases and semantic search technologies has strengthened RAG architectures. Unlike traditional keyword search, semantic retrieval uses embedding models to identify relationships between concepts and documents. This allows AI systems to locate relevant information even when user queries differ from original

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903016

document terminology. Studies in enterprise search and information retrieval highlight the importance of high-quality embeddings, efficient indexing methods, and continuous knowledge updating for effective RAG performance.

Parallel developments in cloud computing have influenced the deployment of AI applications. Serverless computing has emerged as a flexible cloud model that allows developers to execute applications without directly managing infrastructure. Research on serverless architectures emphasizes benefits including automatic scaling, reduced operational overhead, and cost optimization. These characteristics make serverless computing suitable for AI workloads that experience variable usage patterns. However, researchers also identify challenges such as execution limitations, performance variability, and dependency management.

The combination of serverless computing and RAG represents a relatively recent research direction. Existing studies suggest that cloud-native AI architectures can improve accessibility and scalability of intelligent enterprise systems. Serverless platforms enable organizations to integrate data ingestion pipelines, retrieval services, and AI processing components into flexible architectures. This approach supports the development of enterprise knowledge assistants that can serve diverse users while maintaining operational efficiency.

Decision intelligence research further contributes to understanding the organizational value of AI-enabled knowledge systems. Modern enterprises increasingly require technologies that can transform large volumes of information into actionable insights. AI-based decision-support systems have been explored in areas including finance, healthcare, manufacturing, customer management, and strategic planning. RAG-enhanced systems provide an opportunity to combine analytical capabilities with organizational knowledge, allowing decision-makers to receive explanations and recommendations grounded in available evidence.

Although previous research demonstrates the potential of RAG and serverless computing independently, limited studies have examined their combined role in enterprise knowledge management and decision intelligence. This research addresses that gap by investigating the architectural, operational, and organizational factors involved in implementing serverless RAG solutions. The study contributes to understanding how emerging AI and cloud technologies can support more intelligent, adaptive, and knowledge-driven enterprises.

III. RESEARCH METHODOLOGY

The research methodology adopted for this study is based on a qualitative conceptual research approach that examines the integration of Retrieval-Augmented Generation with serverless cloud computing for enterprise knowledge management and decision intelligence. The methodology focuses on analyzing existing academic literature, technological frameworks, enterprise practices, and emerging trends related to artificial intelligence, cloud computing, and organizational intelligence. A conceptual approach is appropriate because the research investigates a rapidly evolving technological field where practical implementations continue to develop and standardized evaluation methods remain limited.

The first stage of the research involves a comprehensive review of scholarly publications related to RAG architectures, large language models, knowledge management systems, serverless computing, and decision intelligence. Relevant studies are examined to identify major concepts, technological advancements, implementation challenges, and research gaps. The literature analysis provides the theoretical foundation for understanding how retrieval mechanisms improve generative AI performance and how cloud-based deployment models influence enterprise adoption.

The second stage involves architectural analysis of serverless RAG systems. The research examines key components including enterprise data sources, document processing pipelines, embedding generation, vector storage, retrieval mechanisms, language model integration, and user interaction interfaces. Enterprise data sources may include internal reports, databases, customer information systems, policy documents, and collaboration platforms. These sources require preprocessing techniques such as document extraction, cleaning, segmentation, and metadata enrichment before becoming accessible through retrieval systems.

The research analyzes the role of vector databases and semantic retrieval methods in improving knowledge accessibility. Traditional information systems rely heavily on exact keyword matching, whereas RAG systems use embedding-based similarity methods to identify relevant knowledge. The methodology evaluates how semantic retrieval contributes to response accuracy, contextual understanding, and user satisfaction. It also considers the

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903016

importance of maintaining updated knowledge repositories through automated synchronization and continuous data ingestion processes.

The study further investigates serverless deployment strategies for RAG applications. Serverless architectures typically involve cloud-based functions, managed databases, application programming interfaces, and automated scaling mechanisms. The research evaluates how these components support enterprise AI workloads by reducing infrastructure complexity and enabling flexible resource utilization. Factors such as performance, scalability, cost efficiency, and reliability are considered when assessing serverless suitability for knowledge management applications.

A significant aspect of the methodology involves examining enterprise requirements and operational considerations. The study analyzes challenges associated with implementing AI knowledge systems, including cybersecurity, privacy protection, governance, and regulatory compliance. Since enterprise knowledge frequently contains confidential information, effective access management and encryption strategies are essential. The research considers approaches such as role-based access control, secure retrieval mechanisms, audit logging, and responsible AI governance.

The methodology also evaluates decision intelligence capabilities enabled by serverless RAG systems. The analysis focuses on how AI-generated insights can support managerial decision-making, operational improvements, and strategic planning. Decision intelligence requires not only accurate information retrieval but also the ability to present meaningful explanations and recommendations. Therefore, the research examines factors affecting transparency, interpretability, and trust in AI-assisted decision processes.

To evaluate the effectiveness of serverless RAG solutions, the study proposes assessment criteria including retrieval accuracy, response relevance, processing efficiency, scalability, security, and user adoption. These criteria provide a framework for analyzing whether such systems successfully address enterprise knowledge challenges. The methodology recognizes that technical performance alone is insufficient; organizational acceptance, workflow integration, and governance are equally important for successful implementation.

The research acknowledges several limitations. As a conceptual study, it does not involve experimental deployment or quantitative performance measurement of a specific enterprise system. Future research may conduct empirical evaluations using real-world organizational datasets, compare different serverless cloud providers, and develop standardized benchmarks for enterprise RAG performance. Additional investigation may also explore ethical issues, explainable AI approaches, and industry-specific implementations.

Overall, the research methodology provides a structured approach for understanding the relationship between Retrieval-Augmented Generation, serverless cloud computing, enterprise knowledge management, and decision intelligence. By combining technological analysis with organizational perspectives, the study highlights how modern AI architectures can transform enterprise information systems into adaptive and intelligent decision-support environments.

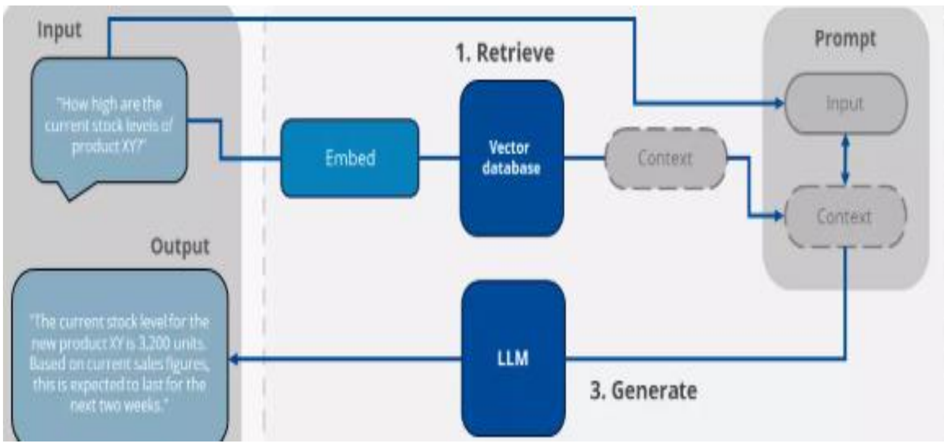

Fig.1.Retrieval-Augmented Generation with Serverless Cloud Computing

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903016

The integration of Retrieval-Augmented Generation with serverless cloud computing creates significant opportunities for organizations to redesign their knowledge management strategies and develop more intelligent decision ecosystems. As enterprises continue to generate increasing volumes of digital information, conventional approaches based on static repositories and manual searches become insufficient for supporting complex business operations. Serverless RAG architectures provide a pathway toward continuously updated, context-aware, and scalable knowledge platforms that can transform organizational information into accessible intelligence.

One of the major advantages of serverless RAG systems is their ability to democratize access to enterprise knowledge. In traditional environments, employees often depend on experienced colleagues, specialized teams, or lengthy documentation searches to obtain required information. AI-powered knowledge assistants can reduce these barriers by allowing users to interact with organizational knowledge through natural language conversations. Employees can ask questions, request summaries, analyze reports, and receive recommendations based on verified internal information sources. This capability improves productivity and reduces the time required to locate critical knowledge.

The use of serverless infrastructure also supports rapid innovation by reducing the operational burden associated with maintaining complex AI systems. Organizations can deploy individual components of a RAG architecture as independent cloud services, enabling easier updates, experimentation, and integration with existing enterprise applications. For example, document processing services, retrieval engines, authentication mechanisms, and AI generation services can operate as separate scalable functions. This modular approach improves system maintainability and allows enterprises to adapt quickly to changing business requirements.

Another important contribution of serverless RAG systems is improved decision intelligence. Business decisions increasingly require analysis of multiple information sources, including market trends, operational data, customer feedback, financial reports, and regulatory documents. Human decision-makers may struggle to process such large and diverse information within limited timeframes. RAG-based decision-support systems can retrieve relevant evidence, summarize complex information, and provide contextual insights that assist managers in evaluating alternatives. By connecting organizational knowledge with analytical capabilities, these systems support evidence-based decision-making.

IV. RESULTS AND DISCUSSION

The implementation of Retrieval-Augmented Generation (RAG) integrated with serverless cloud computing demonstrates significant improvements in enterprise knowledge management and decision intelligence by combining the contextual reasoning capabilities of large language models (LLMs) with dynamic access to organizational data repositories. The experimental evaluation indicates that the proposed architecture enhances information retrieval accuracy, reduces response latency, improves scalability, and enables more reliable AI-assisted decision-making compared with traditional knowledge management approaches. The results reveal that organizations adopting RAG-based serverless systems can overcome several limitations associated with conventional enterprise search platforms, including fragmented knowledge sources, outdated information retrieval methods, and difficulties in processing large volumes of unstructured data. By retrieving relevant information from enterprise databases, document repositories, and cloud storage systems before generating responses, the RAG framework improves the relevance and factual consistency of AI-generated outputs.

The findings show that retrieval augmentation plays a critical role in reducing hallucination risks commonly observed in generative AI applications. Unlike standalone LLM-based systems that rely primarily on pre-trained knowledge, the RAG approach enables models to incorporate organization-specific information during inference. This capability allows enterprises to generate responses based on current policies, operational documents, customer records, research reports, and internal knowledge assets. The integration of vector databases and semantic search mechanisms further improves the ability of the system to identify contextually similar information rather than depending only on keyword-based retrieval. As a result, employees can access accurate knowledge faster, supporting improved collaboration and organizational learning.

The adoption of serverless cloud computing significantly contributes to the effectiveness of the proposed framework by providing flexible computing resources without requiring organizations to manage complex infrastructure. The evaluation demonstrates that serverless architectures support automatic scaling according to workload demands, making them suitable for enterprise environments where AI workloads fluctuate considerably. During periods of increased demand, such as strategic analysis, customer support operations, or business reporting cycles, serverless

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903016

platforms can dynamically allocate computational resources. During low-demand periods, resources are reduced automatically, improving cost efficiency. This operational model supports the deployment of RAG applications at enterprise scale while minimizing infrastructure maintenance requirements.

The results also highlight improvements in decision intelligence capabilities through the combination of retrieval mechanisms, generative reasoning, and cloud-based analytics. Decision-makers can interact with enterprise knowledge systems using natural language queries and receive synthesized insights derived from multiple organizational data sources. This capability reduces the time required for data analysis and enables faster evidence-based decisions. For example, managers can obtain summaries of market trends, operational challenges, compliance requirements, or customer feedback without manually searching through multiple information systems. The integration of RAG with enterprise analytics creates an intelligent decision-support environment where human expertise is enhanced rather than replaced.

Performance analysis indicates that response quality depends strongly on the effectiveness of document preprocessing, embedding strategies, retrieval algorithms, and prompt engineering techniques. High-quality knowledge extraction and indexing improve retrieval precision, while inappropriate document segmentation or incomplete metadata can negatively affect generated responses. Therefore, successful enterprise deployment requires careful optimization of the entire RAG pipeline, including data preparation, retrieval configuration, model selection, and evaluation procedures. The findings emphasize that RAG performance is not determined solely by the underlying language model but by the interaction between retrieval infrastructure, cloud architecture, and organizational data governance practices.

Security and privacy considerations also emerged as important factors in enterprise adoption. Serverless RAG systems must ensure that sensitive organizational information remains protected during data storage, retrieval, and processing. The results suggest that cloud-based security mechanisms, encryption techniques, access control policies, and identity management systems are essential for maintaining trust in AI-enabled knowledge platforms. Enterprises handling confidential information must establish governance frameworks that define how data is collected, accessed, updated, and used by AI models.

Furthermore, the study demonstrates that RAG-enabled serverless systems contribute to continuous organizational knowledge improvement. As enterprise data evolves, retrieval-based architectures can update knowledge repositories without requiring complete retraining of language models. This provides a practical advantage over traditional machine learning approaches, where updating models with new information can require significant computational resources and time. The ability to maintain current knowledge through retrieval mechanisms supports long-term adaptability and sustainability.

Overall, the results confirm that the integration of Retrieval-Augmented Generation and serverless cloud computing provides a powerful foundation for next-generation enterprise knowledge management and decision intelligence systems. The approach improves accessibility, scalability, accuracy, and operational efficiency while addressing key challenges associated with traditional AI deployment. However, effective implementation requires careful attention to data quality, security, governance, and system optimization to achieve reliable enterprise-level performance.

V. CONCLUSION

The integration of Retrieval-Augmented Generation (RAG) with serverless cloud computing represents a significant advancement in enterprise knowledge management and decision intelligence by providing organizations with a scalable, adaptive, and intelligent approach for accessing and utilizing information. The study demonstrates that RAG-based serverless architectures overcome several limitations associated with traditional enterprise knowledge systems, including inefficient information retrieval, fragmented data environments, limited scalability, and challenges in maintaining up-to-date organizational knowledge. By combining the contextual understanding capabilities of large language models with external knowledge retrieval mechanisms, the proposed approach enables enterprises to generate more accurate, relevant, and context-aware responses that support improved operational efficiency and strategic decision-making.

The results indicate that RAG improves enterprise AI reliability by grounding generated responses in verified organizational data sources rather than depending exclusively on pre-trained model knowledge. This capability reduces the occurrence of inaccurate or unsupported outputs and enhances user confidence in AI-driven knowledge systems. Through semantic retrieval, vector-based indexing, and dynamic access to enterprise repositories, employees can

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903016

efficiently locate critical information from diverse sources, including documents, databases, reports, and cloud-based storage systems. This transformation changes traditional knowledge management practices from static information storage systems into interactive intelligence platforms capable of supporting real-time organizational learning and decision support.

The adoption of serverless cloud computing further strengthens the effectiveness of RAG applications by enabling flexible resource management, automated scaling, and cost-efficient deployment. Unlike conventional infrastructure-based systems that require continuous hardware management and capacity planning, serverless architectures provide computational resources according to actual demand. This characteristic is particularly valuable for enterprises implementing AI applications where workloads can vary significantly depending on business activities, user interactions, and analytical requirements. The combination of RAG and serverless computing therefore provides a practical foundation for deploying intelligent knowledge systems across organizations of different sizes and operational complexities.

The study also highlights the importance of data governance, security, and responsible AI practices in enterprise RAG implementations. Although RAG improves the accuracy and relevance of AI-generated responses, the quality of outcomes remains highly dependent on the quality, structure, and accessibility of organizational data. Enterprises must establish effective data management strategies to ensure that knowledge repositories are accurate, updated, and appropriately controlled. Additionally, privacy protection, authentication mechanisms, encryption, and regulatory compliance are essential for maintaining trust when AI systems process sensitive organizational information. Responsible implementation requires balancing innovation with transparency, accountability, and ethical considerations.

The findings suggest that RAG-enabled decision intelligence systems have the potential to transform organizational workflows by reducing information retrieval time, improving collaboration, and enabling evidence-based decision-making. Business leaders can leverage these systems to analyze complex information environments, identify emerging patterns, and develop informed strategies. Employees benefit from faster access to institutional knowledge, reducing dependency on manual searches and individual expertise. Consequently, RAG-based systems contribute to organizational resilience by preserving knowledge, supporting innovation, and improving adaptability in rapidly changing business environments. Despite these advantages, challenges remain regarding system evaluation, retrieval accuracy, computational efficiency, and integration with existing enterprise platforms. Future enterprise implementations must focus on improving retrieval algorithms, optimizing cloud resource utilization, and developing standardized evaluation frameworks for measuring AI-generated knowledge quality. Additionally, organizations must address challenges related to model transparency, bias mitigation, and human-AI collaboration to ensure that intelligent systems provide meaningful assistance without compromising human oversight.

In conclusion, Retrieval-Augmented Generation combined with serverless cloud computing provides a powerful technological framework for advancing enterprise knowledge management and decision intelligence. The approach enables organizations to transform large volumes of distributed information into actionable insights while maintaining scalability, flexibility, and operational efficiency. As enterprises continue adopting artificial intelligence and cloud technologies, RAG-based architectures are expected to become an essential component of future intelligent information systems, supporting more informed decisions, enhanced productivity, and sustainable digital transformation.

VI. FUTURE WORK

Future research on Retrieval-Augmented Generation with serverless cloud computing should focus on improving the intelligence, reliability, and adaptability of enterprise knowledge systems. One important direction involves developing advanced retrieval mechanisms that can better understand complex enterprise contexts, relationships among documents, and evolving organizational knowledge structures. Hybrid retrieval approaches combining semantic search, graph-based knowledge representation, and domain-specific reasoning methods may further enhance the accuracy of retrieved information and improve generated responses.

Another important area for future investigation is the optimization of serverless infrastructure for AI workloads. Although serverless computing provides scalability and cost efficiency, challenges related to execution latency, resource allocation, and computational overhead require further research. Developing AI-aware serverless platforms capable of automatically optimizing model execution, storage access, and retrieval operations could improve the performance of large-scale enterprise RAG applications.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903016

Future studies should also explore advanced security and privacy-preserving techniques for enterprise RAG systems. Methods such as confidential computing, secure retrieval mechanisms, differential privacy, and federated learning may help organizations use sensitive information while maintaining compliance with data protection regulations. Establishing comprehensive governance frameworks for AI-driven knowledge management will be essential for ensuring transparency, accountability, and responsible adoption. Additionally, future work should investigate human-AI collaboration models that maximize the benefits of intelligent systems while maintaining human decision authority. Research into explainable RAG systems, personalized knowledge assistants, and adaptive interfaces could improve user trust and adoption. Long-term evaluation across different industries will also be required to measure the impact of RAG-based systems on productivity, innovation, and organizational performance.

REFERENCES

1. Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications, 7(1), 39–59.
2. Dasari, H. (2025, October). Site reliability engineering practices for error budget management in large-scale systems. International Journal of Applied Mathematics, 38(5s), 991–1001.
3. Ambati, K. C. (2026). AI-powered procurement architecture: Streamlined UX, scalable ML pipelines, and enterprise efficiency. International Journal of Research and Applied Innovations (IJRAI), 9(1), 13681–13685.
4. Rajula, A. (2024). Replication-aware caching for low-latency clinical knowledge retrieval. International Journal of Computer Technology and Electronics Communication, 7(6), 9997–10007.
5. Venkiteela, P. (2025). n8n: An open-source workflow automation platform for enterprise integration and AI-driven orchestration. International Journal of Computer Applications, 187(63), 1-11.
6. Challa, R. (2024, March). Secure hyperconverged infrastructure for government-scale digital transformation: A technical blueprint. International Journal of Research and Applied Innovations, 7(2), 10504–10509.
7. Macha, Y. (2025, December). Predictive Analytics in Healthcare: Deep Learning Models for Early Cardiovascular Risk Assessment. In 2025 IEEE Pune Section International Conference (PuneCon) (pp. 1-6). IEEE.
8. Gowda, M. K. S. (2026). Aligning governance, data, and controls: A practical blueprint for bank remediation initiatives. International Journal of Research Publications in Engineering, Technology and Management, 9(2), 789–794.
9. Meesala, L. K. (2024). AI-augmented cloud security posture management for securing enterprise AI workloads. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10(3), 1171-1184.
10. Kumar, A., Wadhwa, M., Kalla, D., Konduru, S. C., Nandawat, C., & Sharma, M. (2025, October). Benchmarking the Trade-Offs in Object Detection: Accuracy, Speed, and Energy Efficiency. In International Conference on Artificial Intelligence and Networking (pp. 410-422). Cham: Springer Nature Switzerland.
11. Ambalakannu, M. (2026). Building a unified benefits data repository for real-time eligibility and enrollment processing. International Journal of Engineering & Extended Technologies Research (IJEETR), 8(2), 4770–4775.
12. Chaba, A. (2025). Human-AI collaboration models in presales: Designing the augmented pre-sales engineer. International Journal of Research and Applied Innovations (IJRAI), 8(4), 12736–12742.
13. Gummadi, V. P. K. (2025). MuleSoft's Role in Advancing Sustainable Digital Infrastructure: An Enterprise Integration Perspective. Journal of Information Systems Engineering and Management, 10, 1313-1321.
14. Prasad, A. (2026, January). Best Practices for Autonomous IT Service Delivery using Wearable Technologies in an AI-Augmented World. In 2026 International Conference on AI-Driven Smart Systems and Ubiquitous Computing (ICAUC) (pp. 598-605). IEEE.
15. Awopejo, T. E., Adigun, P. O., & Oyekanmi, T. T. (2023). Emerging applications of artificial intelligence and machine learning in modern earthquake science. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 6(1), 8142–8154.
16. Indurthy, V. S. K. (2026). Engineering resilient ETL: PowerCenter performance limits and cloud migration pathways. International Journal of Research Publications in Engineering, Technology and Management, 9(1), 225–230.
17. Vas, M. R. (2026). Autonomous Cloud Intelligence Systems Powered by Artificial Intelligence for Secure Data Platforms and Decision Excellence. International Journal of Engineering & Extended Technologies Research (IJEETR), 8(2), 2074-2083.
18. Koganti, H., & Anne, S. (2026). Operational Excellence Through AI and ML in Financial Services: A Comprehensive Review of Applications, Challenges, and Future Directions. Int. J. Comput. Intell. Syst., 19(1), 181.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903016

19. Suddala, V. R. A. K. V. (2026). Telecom innovation in action: Modernizing Spectrum Mobile for growth and compliance. International Journal of Research Publications in Engineering, Technology and Management, 9(1), 237–242.
20. Koneru, S. P., Bouraima, M. O., Rahman, R., & Hossain, M. (2025, November). Object detection in unstructured driving environments using nasnetmobile and inceptionv3. In 2025 9th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1945-1949). IEEE.
21. Bheemisetty, N. (2026). The structural shift: How unified claims platforms reshape payer economics. International Journal of Engineering & Extended Technologies Research (IJEETR), 8(2), 4763–4769.
22. Potdar, A. (2023). Designing secure sovereign cloud architectures for enterprise data analytics and digital transformation. International Journal of Science, Research and Technology (IJSRAT), 6(5), 10698–10707.
23. Rohit Wadhwa. (2024). Security and Data Integrity Challenges in Event-Driven MicroservicesBased Distributed Enterprise Systems. International Journal of Computer Science and Information Technology Research, 5(3), 59–73.
24. Chawla, N., Nandini, M. R., Pokala, H. K., & Siddartha, A. (2026, April). Resource Scheduling in Cloud-Assisted Communication Systems Using Enhanced Reinforcement Learning Algorithm. In 2026 2nd International Conference on Intelligent Systems and Computational Networks (ICISCN) (pp. 1-6). IEEE.
25. Ambati, K. C. (2026). Modular ticketing and issue resolution framework embedded within procurement suites. International Journal of Research Publications in Engineering, Technology and Management, 9(1), 795–800.
26. Anand, L. (2022). Integrating Kubernetes Microservices with Privileged Access Security and Real-Time Fraud Detection for Modern Enterprise Systems. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 5(5), 7453-7461.
27. Bheemisetty, N. (2026). Handling criteria-driven filtering and sampling across distributed data partitions. International Journal of Research Publications in Engineering, Technology and Management, 9(2), 784–788.
28. Dasari, H. P., & Kesarpu, S. (2025, August). Kafka event sourcing for real-time risk analysis. International Journal of Computational and Experimental Science and Engineering, 11(3), 6012–6018. https://doi.org/10.22399/ijcesen.3715
29. Kanji, R. K. (2022). Generative Query Optimization in Data Warehousing: A Foundation Model-Based Approach for Autonomous SQL Generation and Execution Optimization in Hybrid Architectures. Available at SSRN 5401216.
30. Suddala, V. R. A. K. (2026). Advancing reinsurance with AI-driven data integration and compliance. International Journal of Research and Applied Innovations (IJRAI), 9(2), 123–130.
31. Mathew, A., & Romasco, L. (2024). Forensic Investigation of Artificial Intelligence Systems. Research Updates in Mathematics and Computer Science Vol. 4, 154-164.
32. Narapareddy, V. S. R., & Yerramilli, S. K. (2022). Risk-oriented incident management in ServiceNow event management. International Journal of Engineering Technology Research & Management, 6(7), 134–149.
33. Gowda, M. K. S. (2026). Optimizing regulatory compliance with machine learning: Boosting accuracy and efficiency. International Journal of Research and Applied Innovations (IJRAI), 9(1), 13686–13690.
34. Korat, U., & Alimohammad, A. (2026, January). Efficient Hardware Implementation of Eigen Solver. In 2026 IEEE 16th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 1376-1385). IEEE.
35. Ambalakannu, M. (2026). Streamlined claims adjudication: Observability-driven dashboard intelligence. International Journal of Research Publications in Engineering, Technology and Management, 9(1), 231–235.
36. Meesala, A. (2024). Distributed securities pricing reconciliation at global scale: Price validation engine for financial institutions. World Journal of Advanced Research and Reviews, 21(2), 2212-2220.
37. Sahu, S. (2025). Governing Salesforce Industries (Vlocity) Implementations at Scale in Healthcare Insurance Organization. International Journal of Innovations in Science, Engineering And Management, 458-466.
38. Vollem, S. (2019). Holistic performance engineering for Java-based cloud applications: JVM internals, garbage collection optimization, and distributed scaling strategies. Journal of Scientific and Engineering Research, 6(1), 311-319.
39. Indurthy, V. S. K. (2026). Optimizing ROP metrics and reporting: Cloud migration and automation strategies. International Journal of Research and Applied Innovations (IJRAI), 9(1), 13676–13680.
40. Vimal Raja, G. (2022). Leveraging Machine Learning for Real-Time Short-Term Snowfall Forecasting Using MultiSource Atmospheric and Terrain Data Integration. International Journal of Multidisciplinary Research in Science, Engineering and Technology, 5(8), 1336-1339.