IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12210

Introduction to Vector Databases for Generative AI: Applications, Performance, Future Projections, and Cost Considerations

Satyadhar Joshi

Independent Researcher, BoFA, Jersey City, NJ, USA

Abstract: The rapid advancement of artificial intelligence (AI), particularly in generative models, has led to an exponential increase in the need for efficient handling of high-dimensional vector data. This paper explores the critical role of vector databases in modern AI applications, focusing on their capabilities, use cases, and the challenges they address.Vector databases have emerged as a critical component in the development of generative AI applications. This paper provides a comprehensive review of the role of vector databases in generative AI, focusing on their ability to store, manage, and retrieve high-dimensional vector data efficiently. This paper explores the critical role of vector databases in modern AI applications, focusing on their capabilities, use cases, and the challenges they address. We examine the fundamental limitations of relational databases in handling vector data, contrasting them with specialized vector databases that are optimized for high-dimensional data storage and similarity search. The paper surveys various vector database solutions, including those offered by major cloud providers like Google, AWS, and Microsoft, and highlights their integration with generative AI frameworks such as Lang Chain, Semantic Kernel, and Vertex AI. We also discuss the impact of vector databases on retrieval-augmented generation (RAG) and other AI-driven applications, emphasizing their ability to enhance the accuracy and relevance of large language models (LLMs). Additionally, the paper provides insights into future trends, including scalability improvements, integration with knowledge graphs, and ethical considerations in AI development. By addressing performance, cost, and implementation challenges, this paper aims to provide a comprehensive understanding of how vector databases are shaping the future of generative AI.

Keywords: Vector Databases, Generative AI, Retrieval-Augmented Generation (RAG), High-Dimensional Data, Machine Learning, AI Applications Vector Databases, Large Language Models

I. INTRODUCTION

The surge in AI applications, especially those involving large language models (LLMs), has brought to the forefront the importance of effectively managing and querying high-dimensional vector data. Vector databases have emerged as a pivotal technology, enabling efficient similarity searches and retrieval of relevant information, which is crucial for applications like semantic search, recommendation systems, and generative AI [1], [2].

Recent studies have explored the limitations of traditional relational databases in handling vector data [3]. While generalized databases like PostgreSQL can support vector data, they often exhibit slower performance compared to specialized vector databases. This paper aims to provide a comprehensive overview of vector databases, their significance in contemporary AI, and the landscape of available solutions.

The rapid advancement of generative AI has necessitated the development of specialized data management systems capable of handling high-dimensional vector data. Traditional relational databases, while effective for structured data, face significant challenges when dealing with vector data due to their inherent limitations in indexing and querying high-dimensional spaces [3]. This has led to the emergence of vector databases, which are specifically designed to manage and retrieve vector data efficiently.

In this paper, we review the current state of vector databases and their role in generative AI applications. We begin by discussing the limitations of relational databases in handling vector data, as highlighted in recent studies [3]. We then explore the advantages of specialized vector databases, including their ability to perform similarity searches and support retrieval-augmented generation (RAG) [4]. Vector databases are designed to store, manage, and query high-dimensional vector data. These vectors, often representing embeddings of text, images, or other data types, capture semantic relationships and enable similarity searches [5], [6].

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12210

A. Key Features

- **High-Dimensional Data Storage:** Vector databases efficiently store and index vectors with hundreds or thousands of dimensions.
- **Similarity Search:** They enable fast and accurate similarity searches using metrics like cosine similarity or Euclidean distance.
- **Scalability and Performance:** Designed to handle large volumes of vector data and high query loads.

B. Use Cases

Vector databases find applications in various AI domains, including:

- **Retrieval-Augmented Generation (RAG):** Enhancing LLMs with external knowledge for more accurate and context-aware responses [7], [8].
- **Semantic Search:** Enabling search based on the meaning of queries rather than keywords [9].
- **Recommendation Systems:** Providing personalized recommendations based on user preferences and similarities between items.
- **Image and Video Retrieval:** Searching and retrieving multimedia content based on visual similarity.

II. REVIEW OF VECTOR DATABASES AND GENERATIVE AI SYNERGIES

Generative AI, particularly LLMs, relies heavily on vector databases for efficient retrieval of relevant information. RAG, a technique that combines LLMs with external knowledge, is a prime example of this synergy [10], [11]. Vector databases have emerged as a crucial technology in the realm of Generative AI (GenAI) applications. Traditional databases often struggle with the complex data structures required for AI, but vector databases are specifically designed to store, manage, and rapidly retrieve high-dimensional vector data [6], [26]. These vectors, represented as arrays of numbers, are clustered based on similarity, enabling efficient similarity searches [26]. Vector databases are pivotal in the evolution of generative AI, offering efficient solutions for managing and retrieving complex vector embeddings [11], [27], [28], [29], [30]. As the field advances, continued research and development in vector database technology will be essential for unlocking the full potential of GenAI.

A. Retrieval-Augmented Generation (RAG)

RAG enhances LLMs by retrieving relevant documents or data from a vector database and incorporating them into the model's input. This approach improves the accuracy and reliability of generated responses by grounding them in factual information [12], [13].

B. Vector Databases and LLMs

The integration of vector databases with LLMs enables various advanced applications, such as:

- **Contextual Chatbots:** Providing more accurate and context-aware responses by retrieving relevant information from a vector database.
- **Knowledge-Intensive Tasks:** Assisting in tasks that require access to vast amounts of knowledge, such as question answering and document summarization.
- **Personalized Content Generation:** Generating content tailored to individual preferences and interests.

C. Vector Database Solutions

Several vector database solutions are available, each with its unique features and capabilities. These include:

- **Specialized Vector Databases:** Pinecone [2], Qdrant [14], Milvus [15].
- **Cloud-Based Vector Databases:** Vertex AI Vector Search [16], [17], Azure Cosmos DB [18], AWS vector data stores [19], [20].
- **Integrated Vector Databases:** Databricks Mosaic AI Vector Search [21], Snowflake Cortex [22], IBM watsonx.data [15], Salesforce Data Cloud [23].
- **Vector Database Libraries and Frameworks:** Spring AI [24], Semantic Kernel [25].

D. The Role of Vector Databases in GenAI

Vector databases serve as a bridge between Large Language Models (LLMs) and external information, providing essential capabilities for GenAI systems [31]. They help overcome the limitations of LLMs by allowing them to access and utilize a broader range of data [8]. This is particularly important in Retrieval-Augmented Generation (RAG) architectures, where vector databases facilitate efficient data retrieval [4], [8]. Several platforms and services, including Azure Cosmos DB [18], Google Cloud [16], [32], and Databricks [21], [33], now offer integrated vector database solutions to support these applications.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12210

E. Key Capabilities and Implementations

The ability of vector databases to handle complex data structures has enhanced the performance of AI systems and opened new possibilities in AI applications [6]. Various solutions have been developed, including specialized vector databases and extensions to existing database systems. For instance, IBM watsonx.data offers an integrated vector database built on Milvus [15], while Google Cloud's AlloyDB AI provides a pgvector-compatible PostgreSQL extension [32]. Furthermore, the integration of vector databases with frameworks like LangChain, Vertex AI, and OpenAI enables the grounding of LLMs for GenAI models [34].

F. Challenges and Considerations

Despite the benefits, choosing the right vector database for a GenAI stack requires careful consideration [35]. Factors such as performance optimization and seamless data integration are critical for success [36]. Additionally, research is ongoing to determine the fundamental limitations of supporting vector data management in relational databases like PostgreSQL [3]. Understanding these limitations is crucial for making informed decisions about the appropriate database architecture for specific AI applications.

G. Use Cases and Applications

Vector databases are used in a wide array of applications. This includes but is not limited to semantic search, similarity search, and recommendation engines [37], [38], [39]. They are also used to enhance AI integrations for Semantic Kernel, to improve data retrieval in RAG, and more [4], [25].

H. Limitations of Relational Databases in Handling Vector Data

Relational databases, such as PostgreSQL, have been widely used for managing structured data. However, they exhibit significant limitations when it comes to handling high-dimensional vector data. As demonstrated by [3], relational databases are not optimized for vector data management, leading to slower performance compared to specialized vector databases. This is primarily due to the lack of efficient indexing mechanisms for high-dimensional data and the inability to perform similarity searches effectively.

I. Advantages of Specialized Vector Databases

Specialized vector databases, on the other hand, are designed to address these limitations. They provide efficient indexing and querying mechanisms for high-dimensional vector data, enabling faster and more accurate similarity searches. This makes them particularly well-suited for generative AI applications, where the ability to retrieve similar vectors quickly is crucial [2].

Moreover, vector databases play a critical role in retrieval-augmented generation (RAG), a technique that enhances the accuracy of generative AI models by retrieving relevant information from external sources [7]. By leveraging vector databases, RAG systems can efficiently retrieve and integrate external knowledge, improving the overall performance of generative AI models [4].

J. Use Cases of Vector Databases in Generative AI

Vector databases have found widespread use in various generative AI applications, including large language models (LLMs) and semantic search. For instance, [10] highlights the importance of vector databases in LLMs, where they are used to store and retrieve embeddings that represent the semantic meaning of text. Similarly, [13] discusses the use of vector databases in semantic search, where they enable efficient retrieval of documents based on their semantic similarity to a query.

An electronic copy can be downloaded from the conference website. For questions on paper guidelines, please contact the conference publications committee as indicated on the conference website. Information about final paper submission is available from the conference website.

K. Current State of Tutorials and Lessons

Several resources provide tutorials and lessons on vector databases and their application in Generative AI. These resources range from beginner-friendly guides to more advanced technical documentation.

- Microsoft's Generative AI for Beginners: Offers a series of lessons, including a module specifically on Retrieval-Augmented Generation (RAG) and vector databases [40]. This resource is excellent for those new to the field.
- Azure Cosmos DB Documentation: Provides tutorials on using Azure Cosmos DB as a vector database, covering various domains and scenarios in analytical and generative AI [18].
- Databricks Documentation: Includes learning resources on Mosaic AI Vector Search, Databricks' vector database solution, explaining its functionality and usage [21].
- Spring AI Reference: Offers reference documentation on Vector Databases for the Spring AI framework [24].

These tutorials and lessons provide valuable hands-on experience and insights into the practical application of vector databases in GenAI.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12210

III. FUTURE TRENDS AND POTENTIAL DIRECTIONS: VECTOR DATABASES AND GENAI (2025-2030)

Based on current trends and research directions, several potential developments can be anticipated in the field of vector databases and Generative AI. Furthermore, based on the current trends and advancements highlighted in the references, we can project the following developments in vector databases and their integration with Generative AI (GenAI) over the next five years.

The field of vector databases and generative AI is rapidly evolving, with several key trends expected to shape the landscape in the coming years. These trends are driven by advancements in technology, increasing demand for scalable AI solutions, and the need for more efficient data management systems. Below, we outline the anticipated progress in vector databases and generative AI, with specific predictions for 2025, 2026, 2027, and 2030.

The future of vector databases in generative AI is bright, with significant advancements expected in integration, scalability, knowledge graph integration, and ethical AI. By 2025, we will see deeper integration with AI frameworks, followed by breakthroughs in scalability and distributed architectures by 2026. The integration of vector databases with knowledge graphs will mature by 2027, and by 2030, ethical AI and responsible data management will become central to the field. These trends will drive further innovation and unlock new possibilities for generative AI applications.

- **Integration with Existing Systems**: The integration of vector databases with existing database systems, such as PostgreSQL via extensions like pgvector, is likely to continue. Research is ongoing to determine the fundamental limitations of such integrations [3], which will inform future development efforts.
- **Enhanced Performance Optimization**: As noted by Latimer [36], optimizing performance will be crucial for vector database success in AI. Future developments will likely focus on improving indexing techniques, search algorithms, and data integration methods to meet the demands of increasingly complex AI applications.
- **Expansion of Use Cases**: Vector databases are currently being applied in various domains, including semantic search and recommendation engines [37], [38]. The range of applications is expected to expand as the technology matures and becomes more accessible [39].

These trends suggest a future where vector databases become even more integral to the GenAI landscape, driving innovation and enabling new possibilities in AI applications.

A. 2025: Enhanced Cloud Integrations with AI Frameworks and Standardization

In 2025, we anticipate significant advancements in cloud-based vector database offerings. Cloud providers like Google Cloud (Vertex AI Vector Search) [16], [17], AWS [19], [20], and Azure (Azure Cosmos DB) [18] will likely introduce more refined and optimized vector database services, focusing on improved scalability and performance. We expect to see greater standardization in APIs and interfaces, simplifying integration with GenAI frameworks. Additionally, the development of more robust data management tools and workflows within these cloud platforms will be a priority.

By 2025, we expect to see deeper integration between vector databases and AI frameworks, particularly in the context of retrieval-augmented generation (RAG) and large language models (LLMs). As highlighted by [16], Google's Vertex AI is already making strides in integrating vector databases with generative AI workflows. By 2025, this trend is likely to expand, with more cloud providers offering native support for vector databases in their AI platforms. Additionally, frameworks like LangChain and Semantic Kernel are expected to further streamline the integration of vector databases with generative AI applications [25].

B. 2026: Hybrid Solutions, Semantic Enrichment, Scalability and Distributed Architecture

By 2026, the focus will shift towards developing hybrid solutions that combine the strengths of vector databases with other technologies like knowledge graphs [12]. This will enable more nuanced and contextually rich information retrieval, improving the accuracy and relevance of GenAI outputs. We expect to see advancements in semantic embedding techniques, allowing vector databases to capture more complex relationships and nuances in data. This will lead to more sophisticated applications in areas like personalized content generation and knowledge-intensive tasks.

Scalability will remain a critical challenge for vector databases, especially as generative AI applications continue to grow in complexity and data volume. By 2026, we anticipate significant progress in distributed architectures for vector databases, enabling them to handle larger datasets and more complex queries efficiently. For instance, [21] discusses Databricks' Mosaic AI Vector Search, which is expected to evolve into a more robust and scalable solution by 2026. This will involve advancements in distributed indexing, parallel processing, and hardware acceleration, making vector databases more suitable for enterprise-level AI applications.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12210

C. 2027: Edge Real-Time Vector Processing and Integration with Knowledge Graphs

In 2027, the emphasis will move towards enabling vector database processing at the edge and in real-time. This will be driven by the increasing demand for low-latency applications, such as real-time chatbots and personalized recommendations. We will likely see the development of lightweight and efficient vector database solutions that can be deployed on edge devices and embedded systems. This will also involve advancements in hardware acceleration and distributed processing techniques.

By 2027, the integration of vector databases with knowledge graphs is expected to become a major trend. While vector databases excel at handling unstructured data, knowledge graphs provide a structured representation of relationships between entities. Combining these two technologies could enable more sophisticated AI applications, such as enhanced semantic search and reasoning over complex datasets. As noted by [12], this integration is still in its early stages, but by 2027, we expect to see more mature solutions that leverage the strengths of both vector databases and knowledge graphs. This will open up new possibilities for generative AI, particularly in domains like healthcare, finance, and e-commerce.

D. 2030: AI-Driven Optimization and Autonomous Management

By 2030, we can anticipate the emergence of AI-driven optimization and autonomous management of vector databases. AI algorithms will be used to automatically tune and optimize vector database performance, adapting to changing workloads and data patterns. Self-managing vector databases will be able to autonomously handle tasks like data indexing, updates, and maintenance. This will significantly reduce the operational overhead and enable developers to focus on building innovative GenAI applications.

By 2030, ethical considerations and responsible data management will become central to the development and deployment of vector databases in generative AI. As vector databases are increasingly used to store sensitive data, such as embeddings derived from personal or proprietary information, there will be a growing need for robust security mechanisms and ethical guidelines. [14] highlights the importance of ethical AI implementation, and by 2030, we expect to see comprehensive frameworks and regulations governing the use of vector databases in AI. This will include advancements in data privacy, bias mitigation, and fairness, ensuring that vector databases are used responsibly in generative AI applications.

E. Broader Ecosystem Integration and Accessibility

Throughout this period, we will see a broader integration of vector databases into the wider AI ecosystem. Tools and frameworks like Spring AI [24], Semantic Kernel [25], and others will provide seamless support for vector database integration, making it easier for developers to leverage these technologies. Additionally, educational resources and training programs, such as those offered by Coursera [41], will contribute to the widespread adoption and accessibility of vector database expertise. These projections are based on the current trajectory and trends reflected in the provided references. However, the field of AI and vector databases is rapidly evolving, and unforeseen innovations may alter these predictions.

IV. VECTOR DATABASE SOLUTIONS BY PROVIDER, INDUSTRY CONTRIBUTIONS: GOOGLE, AWS, AND MICROSOFT

Various cloud providers offer vector database solutions, each with its unique features and integrations. Here's a breakdown of solutions from Google Cloud, Microsoft Azure, and Amazon Web Services (AWS). The rapid adoption of vector databases in generative AI has been significantly influenced by major cloud providers such as Google, AWS, and Microsoft. These companies have developed specialized tools and services to support vector data management, enabling businesses to leverage generative AI more effectively. Below, we discuss the contributions of each provider, as highlighted in the literature. Google, AWS, and Microsoft have each made significant contributions to the development and adoption of vector databases in generative AI. Google's Vertex AI and AlloyDB AI, AWS's vector data stores and Aurora, and Microsoft's Azure Cosmos DB and Semantic Kernel represent key advancements in the field. These providers are driving innovation by making vector databases more accessible, scalable, and integrated with generative AI workflows, paving the way for future advancements in the industry.

A. Google Cloud

Google Cloud provides several options for vector databases, tightly integrated with its AI and data analytics services:

- **AlloyDB AI**: This offering leverages a pgvector-compatible PostgreSQL extension, providing vector search capabilities within a familiar relational database environment [32]. This allows users to combine traditional data management with vector similarity search.
- **Vertex AI RAG Engine**: Google Cloud's Vertex AI offers choices for vector databases within its Retrieval-Augmented Generation (RAG) Engine, facilitating the grounding of LLMs with external knowledge [16].

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12210

B. Microsoft Azure

Microsoft Azure offers Azure Cosmos DB as a fully managed vector database solution:

- **Azure Cosmos DB**: This service enables the use of Azure Cosmos DB as a vector database across various domains and scenarios in analytical and generative AI [18]. Its integration streamlines the development of AI-powered applications. Semantic Kernel is supported via AI integrations [25].

C. Amazon Web Services (AWS)

AWS provides multiple avenues for leveraging vector databases in AI applications:

- **Amazon Aurora**: Amazon Aurora features as a key service for vector data stores in GenAI applications [19], [20].

D. Google: Vertex AI and AlloyDB AI

Google has been at the forefront of integrating vector databases with generative AI workflows. Their Vertex AI platform offers robust support for vector search, enabling efficient retrieval of high-dimensional data for applications like retrieval-augmented generation (RAG) and large language models (LLMs) [16]. Additionally, Google has introduced AlloyDB AI, a PostgreSQL-compatible extension that supports vector search and integrates seamlessly with generative AI frameworks like LangChain [32]. These tools demonstrate Google's commitment to making vector databases more accessible and scalable for AI-driven applications.

E. AWS: Vector Data Stores and Aurora

AWS has also made significant strides in supporting vector databases for generative AI. Their focus has been on integrating vector data stores with existing database solutions like Amazon Aurora. AWS emphasizes the importance of vector data stores in overcoming the limitations of large language models (LLMs) through techniques like retrieval-augmented generation (RAG) [19]. Furthermore, AWS has highlighted the role of vector databases in modern generative AI applications, particularly in enhancing the accuracy and reliability of AI models by retrieving relevant information from external sources [20]. These advancements position AWS as a key player in the vector database ecosystem.

F. Microsoft: Azure Cosmos DB and Semantic Kernel

Microsoft has contributed to the field of vector databases through its Azure Cosmos DB, which now includes integrated vector database capabilities. This allows users to perform vector searches and manage high-dimensional data efficiently, particularly in analytical and generative AI applications [18]. Additionally, Microsoft's Semantic Kernel framework provides integrations with AI services, enabling developers to build generative AI applications that leverage vector databases for tasks like semantic search and retrieval-augmented generation (RAG) [25]. These tools underscore Microsoft's focus on making vector databases a core component of their AI offerings.

V. GAP ANALYSIS AND CONSIDERATIONS: CHALLENGES AND OPPORTUNITIES IN VECTOR DATABASE INTEGRATION WITH GENAI

Choosing the right vector database for a GenAI stack requires careful consideration to ensure it aligns with specific application requirements and technical capabilities. While the integration of vector databases with Generative AI (GenAI) has shown significant promise, several gaps and challenges remain that warrant further attention. This gap analysis identifies key areas where improvements and innovations are needed to fully realize the potential of this synergy. Despite the significant advancements in vector databases and their applications in generative AI, several gaps remain in the current research and technological landscape. These gaps highlight areas where further investigation and innovation are needed to fully realize the potential of vector databases in AI-driven applications.

Based on the literature, several gaps and considerations emerge:

- **Performance Optimization**: Latimer [36] highlights that performance optimization is crucial for vector database success in AI. A gap exists in readily available best practices and tooling for achieving optimal performance in diverse GenAI workloads. More research and community knowledge-sharing are needed in this area.
- **Seamless Data Integration**: Srivastava [35] emphasizes the importance of data integration when selecting a vector database for a GenAI stack. A gap often exists between the theoretical capabilities of a database and its practical ability to integrate smoothly with existing data pipelines and AI frameworks.
- **Limitations of Relational Database Extensions**: Zhang et al. [3] investigate the fundamental limitations of supporting vector data management in relational databases like PostgreSQL. Understanding these limitations is crucial to avoid potential performance bottlenecks and scalability issues when relying on extensions like pgvector. This reveals a gap in fully understanding the trade-offs between specialized vector databases and relational database extensions for vector data.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12210

- **Evaluation and Selection Criteria**: A clear framework for evaluating and selecting vector databases is currently lacking. While resources like Srivastava's article [35] provide guidance, a more comprehensive and standardized set of criteria is needed to help organizations make informed decisions.

Addressing these gaps through further research, development, and community collaboration will be essential for the continued advancement and effective adoption of vector databases in GenAI applications.

A. Scalability and Performance Limitations

One of the primary challenges is the scalability and performance of vector databases when dealing with extremely large datasets and high query loads. As the volume of vector data grows, maintaining low latency and high throughput becomes increasingly difficult. Current solutions may struggle to handle the demands of real-time applications and large-scale deployments [9]. There is a need for more efficient indexing and search algorithms that can scale effectively with increasing data volumes.

B. Data Management and Update Challenges

Efficiently managing and updating vector data is another critical challenge. In dynamic environments where data is constantly changing, maintaining the accuracy and relevance of vector embeddings requires frequent updates. Current systems may lack robust mechanisms for incremental updates and data versioning, leading to inconsistencies and stale information. Developing solutions that support efficient data management and updates is crucial for maintaining the quality of GenAI applications [35].

C. Integration Complexity and Standardization

The integration of vector databases with various GenAI frameworks and tools can be complex and time-consuming. Lack of standardization across different vector database solutions and AI platforms creates interoperability challenges. There is a need for standardized APIs and interfaces that simplify the integration process and enable seamless data exchange between different systems. Additionally, better tooling and documentation are needed to assist developers in effectively integrating vector databases into their GenAI workflows [24], [25].

D. Choice and Optimization of Vector Databases

Selecting the right vector database for a specific GenAI application can be challenging. Different vector databases offer varying features, performance characteristics, and trade-offs. There is a need for clear guidelines and benchmarks that help developers choose the most suitable solution for their specific needs. Furthermore, optimizing vector database performance for specific GenAI workloads requires expertise and experimentation. Developing best practices and tools for performance optimization is essential for maximizing the efficiency of GenAI applications.

E. Semantic Understanding and Contextual Relevance

While vector databases excel at similarity search, ensuring semantic understanding and contextual relevance remains a challenge. Current methods may struggle to capture complex relationships and nuances in the data. There is a need for more advanced embedding techniques and search algorithms that can better capture the semantic meaning of data and provide more contextually relevant results. Research into hybrid approaches that combine vector search with knowledge graphs and other semantic technologies may offer promising solutions [12].

F. Security and Privacy Considerations

As vector databases store sensitive information, ensuring security and privacy is paramount. Developing robust security mechanisms, such as access control, encryption, and data anonymization, is crucial for protecting sensitive data. Additionally, addressing privacy concerns related to the use of vector databases in GenAI applications requires careful consideration of data governance and compliance requirements. By addressing these gaps and challenges, we can unlock the full potential of vector database integration with GenAI and drive innovation across various domains.

G. Limitations in Relational Databases

One of the primary gaps lies in the limitations of relational databases when handling high-dimensional vector data. As highlighted by [3], relational databases like PostgreSQL are not inherently designed to manage vector data efficiently. While extensions such as pgvector have been developed to support vector operations, these solutions often fall short in terms of performance and scalability compared to specialized vector databases. This gap underscores the need for further research into optimizing relational databases for vector data management or developing hybrid solutions that combine the strengths of both relational and vector databases.

H. Scalability and Performance

Another critical gap is the scalability and performance of vector databases in large-scale generative AI applications. While vector databases excel in similarity searches and retrieval-augmented generation (RAG) [4], their performance

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12210

can degrade as the volume of data increases. This is particularly relevant in applications involving large language models (LLMs), where the number of embeddings and the complexity of queries can grow exponentially. Research into more efficient indexing algorithms, distributed storage solutions, and hardware acceleration techniques is needed to address these scalability challenges [10].

I. Integration with Knowledge Graphs

A promising yet underexplored area is the integration of vector databases with knowledge graphs. While vector databases excel at handling unstructured data, knowledge graphs provide a structured representation of relationships between entities. Combining these two technologies could enable more sophisticated AI applications, such as enhanced semantic search and reasoning over complex datasets. However, as noted by [12], there is limited research on how to effectively integrate vector databases with knowledge graphs, particularly in the context of generative AI. This gap presents an opportunity for future research to explore hybrid architectures that leverage the strengths of both technologies.

J. Ethical and Security Concerns

Finally, there is a notable gap in addressing the ethical and security concerns associated with vector databases in generative AI. As vector databases are increasingly used to store sensitive data, such as embeddings derived from personal or proprietary information, there is a growing need for robust security mechanisms to prevent unauthorized access and data breaches. Additionally, the use of vector databases in AI applications raises ethical questions about data privacy, bias, and fairness. While some work has been done in this area [14], more comprehensive frameworks and guidelines are needed to ensure the responsible use of vector databases in AI.

K. Conclusion in Gaps Discussed

In summary, while vector databases have made significant strides in supporting generative AI applications, several gaps remain in terms of relational database integration, scalability, knowledge graph integration, and ethical considerations. Addressing these gaps will require interdisciplinary research and collaboration across the fields of database systems, AI, and cybersecurity. By closing these gaps, the research community can unlock the full potential of vector databases and drive further innovation in generative AI.

VI. AI FRAMEWORKS AND VECTOR DATABASE INTEGRATION

Several AI frameworks are increasingly leveraging vector databases to enhance their capabilities, particularly in Generative AI applications. This integration enables more efficient data retrieval, improved model grounding, and enhanced overall performance. The integration of vector databases with AI frameworks has become a critical enabler for generative AI applications. Several AI frameworks have been developed to facilitate the use of vector databases in tasks such as retrieval-augmented generation (RAG), semantic search, and large language model (LLM) workflows. Below, we discuss the AI frameworks mentioned in the literature and their association with vector databases.

AI frameworks like LangChain, Semantic Kernel, and Vertex AI play a crucial role in bridging the gap between vector databases and generative AI applications. These frameworks provide the tools and integrations necessary to leverage vector databases for tasks such as semantic search, retrieval-augmented generation (RAG), and large language model (LLM) workflows. By enabling efficient management and retrieval of high-dimensional data, these frameworks are driving innovation in the field of generative AI and expanding the use cases for vector databases.

The integration of vector databases has become a cornerstone in the development of robust Generative AI (GenAI) products. This synergy addresses the inherent limitations of Large Language Models (LLMs) by providing access to external, up-to-date information, enhancing accuracy and contextuality.

- **LangChain**: The integration of vector databases with frameworks like LangChain enables the grounding of Large Language Models (LLMs) for Generative AI models [34]. This allows LLMs to access and utilize external knowledge stored in vector databases, improving the accuracy and relevance of their responses.
- **Vertex AI**: Google Cloud's Vertex AI offers vector database choices within its Retrieval-Augmented Generation (RAG) Engine [16]. This integration facilitates the development of GenAI applications by providing tools and services for managing vector embeddings and performing similarity searches.
- **Semantic Kernel**: Microsoft's Semantic Kernel is supported via AI integrations with Azure Cosmos DB [25]. This simplifies the development of intelligent applications by providing a unified platform for AI development.

These integrations demonstrate the growing importance of vector databases in the AI ecosystem, providing developers with the tools and infrastructure needed to build powerful and intelligent applications.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12210

A. LangChain

LangChain is a prominent AI framework that has gained attention for its ability to integrate vector databases with generative AI workflows. It provides tools for building applications that leverage retrieval-augmented generation (RAG), enabling AI models to retrieve relevant information from external sources using vector databases. As highlighted by [32], LangChain's compatibility with vector databases like Google's AlloyDB AI and pgvector extensions has made it a popular choice for developers working on generative AI applications. This integration allows for efficient semantic search and retrieval, enhancing the accuracy and relevance of AI-generated outputs.

B. Semantic Kernel

Microsoft's Semantic Kernel is another AI framework that supports the integration of vector databases into generative AI applications. It provides a unified platform for building AI-driven solutions, including those that require vector search and retrieval capabilities. According to [25], Semantic Kernel offers integrations with AI services and vector databases, enabling developers to create sophisticated applications that leverage high-dimensional data for tasks like semantic search and RAG. This framework is particularly useful for enterprises looking to incorporate vector databases into their AI ecosystems.

C. Vertex AI

Google's Vertex AI is a comprehensive AI platform that supports vector databases as part of its generative AI capabilities. Vertex AI enables users to perform vector searches and manage high-dimensional data efficiently, making it a key tool for applications like retrieval-augmented generation (RAG) and large language models (LLMs). As noted by [16], Vertex AI's integration with vector databases allows for seamless retrieval of relevant information, improving the performance of generative AI models. This framework is particularly well-suited for businesses leveraging cloud-based AI solutions.

D. Enhancing LLMs with Vector Databases

Retrieval-Augmented Generation (RAG) is a key technique that leverages vector databases to augment LLMs. By retrieving relevant documents or data, RAG allows LLMs to generate more accurate and context-aware responses [7]. Vector databases store embeddings of documents, enabling efficient similarity searches that provide LLMs with the necessary context [8], [10].

E. Product Integrations and Use Cases

Several GenAI products and platforms have integrated vector databases to enhance their capabilities. For instance, cloud providers like Google Cloud and AWS offer vector database services, such as Vertex AI Vector Search [16], [17] and AWS vector data stores [19], [20], enabling developers to build powerful GenAI applications. Similarly, integrated solutions like Databricks Mosaic AI Vector Search [21], Snowflake Cortex [22], and IBM watsonx.data [15] provide vector database capabilities within their platforms.

Use cases span various domains, including contextual chatbots, knowledge-intensive tasks, and personalized content generation. In these scenarios, vector databases facilitate the retrieval of relevant information, ensuring that GenAI products deliver accurate and contextually appropriate outputs [12], [13].

F. Vector Databases as a Core Component

The role of vector databases is not limited to RAG. They serve as a fundamental component in building enterprise-ready GenAI applications. By providing efficient similarity search and retrieval, vector databases enable GenAI products to access and utilize vast amounts of data, addressing the challenge of keeping LLMs up-to-date and contextually relevant [11]. This integration ensures that GenAI products can deliver more reliable and accurate results, driving innovation across various industries.

VII. QUANTITATIVE PERFORMANCE AND COST CONSIDERATIONS

When evaluating vector database solutions for Generative AI applications, quantitative performance metrics and cost considerations are paramount. These factors directly impact the efficiency, scalability, and overall return on investment of AI initiatives.

The adoption of vector databases in generative AI applications brings with it important considerations regarding performance and cost. While vector databases offer significant advantages in terms of efficiency and scalability, their implementation must be carefully evaluated to ensure optimal performance and cost-effectiveness. Below, we discuss key findings related to the quantitative performance and cost considerations of vector databases, as highlighted in the literature. Quantitative performance and cost considerations are critical when adopting vector databases for generative AI applications. While specialized vector databases offer superior performance for high-dimensional data, their implementation and maintenance can be costly.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12210

Relational databases, though less expensive, may not provide the same level of efficiency. Therefore, organizations must carefully evaluate their specific needs and explore hybrid or optimized solutions to achieve a balance between performance and cost. These considerations will play a key role in the widespread adoption of vector databases in generative AI.

The adoption of vector databases in Generative AI applications involves significant financial and cost considerations that organizations must carefully evaluate.

- **Total Cost of Ownership (TCO)**: Choosing a vector database for a GenAI stack involves assessing the total cost of ownership, encompassing infrastructure, software licensing, and operational expenses [35]. Selecting a cost-effective solution is essential.

A thorough cost-benefit analysis is crucial to justify the investment in vector database technology for GenAI initiatives.

A. Contextual Numerical Significance of Vector Databases

While the provided bibliography lacks direct quantitative performance benchmarks, we can infer the numerical significance of vector databases through contextual indicators of their adoption and relevance in the rapidly expanding Generative AI (GenAI) landscape.

- **Performance Metrics**: Key performance indicators (KPIs) for vector databases include query latency, throughput (queries per second), and indexing speed. Achieving low query latency is critical for real-time applications, while high throughput ensures the system can handle a large volume of requests [36]. Indexing speed affects the time required to ingest and prepare data for search.
- **Scalability**: The ability to scale horizontally to accommodate growing data volumes and user demands is crucial. Vector databases should be able to maintain consistent performance as the dataset size and query load increase. Scalability directly impacts infrastructure costs and operational efficiency.
- **Cost Optimization**: The total cost of ownership (TCO) for a vector database solution includes infrastructure costs (e.g., compute, storage, networking), software licensing fees, and operational expenses. Selecting a cost-effective solution that aligns with specific performance requirements is essential [35].

Quantitative analysis of these performance and cost factors is essential for making informed decisions about vector database adoption in GenAI deployments.

B. Performance of Vector Databases

Vector databases are designed to handle high-dimensional data efficiently, enabling fast similarity searches and retrieval operations. However, their performance can vary significantly depending on the implementation and the scale of the data. For instance, [3] conducted a case study on PostgreSQL's ability to manage vector data and found that relational databases exhibit slower performance compared to specialized vector databases. This performance gap is primarily due to the lack of optimized indexing mechanisms for high-dimensional data in relational systems. Specialized vector databases, on the other hand, are designed to perform these operations more efficiently, making them better suited for generative AI applications that require real-time retrieval of similar vectors.

C. Cost Implications

The cost of implementing and maintaining vector databases is another critical factor to consider. While specialized vector databases offer superior performance, they often come with higher infrastructure and operational costs. For example, cloud-based vector database solutions, such as those offered by Google's Vertex AI [16] and AWS's vector data stores [19], provide scalable and efficient solutions but may incur significant costs depending on the volume of data and the frequency of queries. Additionally, the integration of vector databases with AI frameworks like LangChain and Semantic Kernel [25] may require additional resources for development and maintenance, further impacting the overall cost.

D. Balancing Performance and Cost

To achieve a balance between performance and cost, organizations must carefully evaluate their specific use cases and requirements. For instance, [10] suggests that hybrid approaches, combining the strengths of relational and vector databases, may offer a cost-effective solution for certain applications.

Similarly, [21] highlights the importance of optimizing indexing and querying mechanisms to reduce computational overhead and associated costs. By leveraging these strategies, organizations can maximize the performance of vector databases while minimizing their operational expenses.

E. Exponential Growth of Vector Data

The rise of LLMs and embedding-based applications has led to an exponential increase in the volume of high-dimensional vector data. Though not explicitly stated, the sheer number of companies integrating vector databases, as

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12210

evidenced by references to platforms like Vertex AI [16], [17], Azure Cosmos DB [18], Databricks Mosaic AI [21], and others, implies massive data growth. Each of these platforms is designed to handle very large datasets.
Implied Scale: The fact that major cloud providers are investing heavily in vector database solutions suggests that the volume of vector data being generated and stored is in the petabyte to exabyte scale.

F. Broad Adoption in AI Applications

The references highlight the widespread adoption of vector databases across various AI applications, including RAG [7], [8], [10], [12], [13], semantic search [9], and recommendation systems.
Implied Usage: The number of applications relying on semantic search and RAG is rapidly increasing, which indicates that millions, if not billions, of similarity searches are performed daily.

G. Cloud Provider Investment

The fact that cloud providers like AWS [19], [20], Google Cloud [16], [17], and Microsoft Azure [18] are heavily investing in vector database services indicates the financial and technological significance of this technology.
Implied Investment: Cloud providers invest large amounts of money in their infrastructure and services. The fact that they have implemented vector database services indicates that this is a very important and profitable endeavor.

H. Industry-Wide Adoption

The references show that vector databases are being adopted across various industries.
Implied Usage: The adoption of vector databases across so many industries implies that a very large amount of data is being processed, and that the amount of users of vector based systems is growing quickly.

I. Focus on Real-Time Applications

The references suggest a growing emphasis on real-time applications, such as chatbots and personalized recommendations, which require low-latency vector search.
Implied Performance: The demand for real-time applications indicates that vector databases are being designed to achieve sub-second query latencies, which is a significant performance requirement.

J. Cost Considerations in Vector Database Adoption: Inferences

While the literature does not provide explicit financial figures or dollar amounts, several references highlight the importance of cost considerations when adopting vector databases for generative AI applications. These considerations include infrastructure costs, operational expenses, and the trade-offs between performance and affordability. Below, we discuss the implied financial implications of using vector databases, as derived from the available literature.

K. Infrastructure Costs

The adoption of specialized vector databases often requires significant infrastructure investments. For example, cloud-based solutions like Google's Vertex AI [16] and AWS's vector data stores [19] offer scalable and efficient platforms for managing high-dimensional data. However, these services typically operate on a pay-as-you-go model, where costs scale with data volume, query frequency, and computational resources. While no specific dollar amounts are provided in the literature, it is clear that organizations must budget for these variable costs when deploying vector databases in large-scale generative AI applications.

L. Operational Expenses

Operational expenses are another critical factor in the adoption of vector databases. Integrating vector databases with AI frameworks like LangChain and Semantic Kernel [25] may require additional development and maintenance efforts, which can increase overall costs. Furthermore, specialized vector databases often demand skilled personnel for setup, optimization, and ongoing management, adding to the operational burden. Although the literature does not quantify these expenses, it emphasizes the need for organizations to account for these hidden costs when planning their AI infrastructure.

M. Performance vs. Cost Trade-offs

The trade-off between performance and cost is a recurring theme in the literature. For instance, [3] highlights the performance limitations of relational databases like PostgreSQL when handling vector data, suggesting that specialized vector databases, while more expensive, offer superior efficiency. On the other hand, hybrid approaches that combine relational and vector databases may provide a more cost-effective solution for certain use cases [10]. Organizations must carefully evaluate their specific requirements to determine whether the performance benefits of specialized vector databases justify the associated costs.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12210

N. Implied Financial Significance of Vector Databases

While direct financial figures are absent from the provided bibliography, the references strongly suggest a substantial financial impact and market potential for vector database technologies.

O. Significant Cloud Provider Investment

The heavy investment by major cloud providers—AWS [19], [20], Google Cloud [16], [17], and Microsoft Azure [18]—in vector database services indicates a substantial financial commitment. Cloud providers allocate significant resources to develop and deploy services they anticipate will generate substantial revenue.

Implied Market Size: The fact that these major players are investing heavily suggests a multi-billion dollar market opportunity for vector database solutions. Implied Revenue Streams: These services are integrated into broader cloud offerings, implying that they contribute to substantial revenue streams through subscriptions and usage-based billing.

P. Integration into Enterprise-Level Platforms

The integration of vector databases into enterprise-level platforms like Databricks Mosaic AI [21], Snowflake Cortex [22], and IBM watsonx.data [15] further underscores their financial significance. These platforms cater to large enterprises with substantial budgets for data analytics and AI solutions.

Implied Enterprise Spending: The inclusion of vector databases in these platforms indicates that enterprises are willing to invest significant sums in these technologies to enhance their data-driven capabilities. Implied Increased Value: The addition of vector database functionality increases the value proposition of these existing platforms, which leads to increased revenue for these companies.

Q. Driving Innovation in GenAI Applications

Vector databases are crucial for enabling advanced GenAI applications, which are rapidly transforming various industries. This transformation implies significant financial benefits for companies that leverage these technologies.

Implied ROI: Improved accuracy and efficiency in AI applications, such as RAG [7], [8], [10], [12], [13] and semantic search [9], can lead to substantial cost savings and revenue growth. Implied Market Expansion: The ability to develop innovative GenAI products and services using vector databases creates new market opportunities and drives further investment in this technology.

R. Increased Demand for Specialized Skills

The growing adoption of vector databases creates demand for specialized skills and expertise, which translates into increased investment in training and education.

Implied Investment in Talent: Companies will invest in hiring and training professionals with expertise in vector database technologies, leading to increased salaries and market demand for these skills. Implied Educational Market: The growth of educational resources and training programs, such as those offered by Coursera [41], indicates a growing market for vector database education.

S. Final words on Cost Analysis

While the literature does not provide specific financial numbers or dollar amounts, it underscores the importance of considering infrastructure costs, operational expenses, and performance trade-offs when adopting vector databases for generative AI applications. Organizations must carefully evaluate their budgets and requirements to ensure that the benefits of vector databases align with their financial constraints. Future research that quantifies these costs in detail would provide valuable insights for businesses planning to invest in vector database technologies.

VIII. CHALLENGES, FUTURE DIRECTIONS AND CONCLUSION

Despite their advancements, vector databases face several challenges, including:

- **Scalability and Performance:** Handling extremely large datasets and high query loads.
- **Data Management:** Efficiently managing and updating vector data.
- **Integration with AI Frameworks:** Seamlessly integrating with LLMs and other AI tools.
- **Choice and Implementation:** Selecting the right vector database for a specific use case [9], [35].

Future research directions include exploring hybrid approaches that combine the strengths of specialized and generalized databases, optimizing vector search algorithms, and developing more robust and scalable vector database solutions. As generative AI continues to evolve, the role of vector databases is expected to grow. Future research should focus on improving the scalability and performance of vector databases, particularly in the context of large-scale generative AI applications. Additionally, the integration of vector databases with other AI technologies, such as knowledge graphs, presents an exciting opportunity for further innovation [12].

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 2, February 2025

DOI: 10.17148/IARJSET.2025.12210

Vector databases have become an indispensable technology in the era of AI, particularly in generative models. Their ability to efficiently manage and query high-dimensional vector data enables various applications, from RAG to semantic search. As AI continues to evolve, the importance of vector databases will only grow, driving further innovation and research in this field. In conclusion, vector databases have become an indispensable tool in the development of generative AI applications. Their ability to efficiently store, manage, and retrieve high-dimensional vector data makes them well-suited for a wide range of use cases, including retrieval-augmented generation and large language models. As the field of generative AI continues to advance, the importance of vector databases is likely to increase, driving further innovation in this area.

REFERENCES

[1]. "Vector Databases: A Technical Primer." Accessed: Feb. 22, 2025. [Online] https://medium.com/@babajide.ogunjobi/vector-databases-a-technical-primer-84cbe42885ac
[2]. "What is a Vector Database & How Does it Work? Use Cases + Examples Pinecone." Accessed: Feb. 22, 2025. [Online]. Available: https://www.pinecone.io/learn/vector-database/
[3]. Y. Zhang, S. Liu, and J. Wang, "Are There Fundamental Limitations in Supporting Vector Data Management in Relational Databases? A Case Study of PostgreSQL," in 2024 IEEE 40th International Conference on Data Engineering (ICDE), Utrecht, Netherlands: IEEE, May 2024, pp. 3640–3653. doi: 10.1109/ICDE60146.2024.00280.
[4]. "Vector Databases for Efficient Data Retrieval in RAG: Unlocking the Future of AI and Recruitment Technology Medium." Accessed: Feb. 22, 2025. [Online]. Available: https://medium.com/@genuine.opinion/vector-databases-for-efficient-data-retrieval-in-rag-a-comprehensive-guide-dcfcbfb3aa5d
[5]. "What is a Vector Database & How Does it Work? Use Cases + Examples Pinecone." Accessed: Feb. 24, 2025. [Online]. Available: https://www.pinecone.io/learn/vector-database/
[6]. "Vector Database used in AI Exxact Blog." Accessed: Feb. 24, 2025. [Online]. Available: https://www.exxactcorp.com/blog/deep-learning/vector-database-for-llms-generative-ai-and-deep-learning
[7]. R. Merritt, "What Is Retrieval-Augmented Generation aka RAG?" NVIDIA Blog. Jan. 2025. Accessed: Feb. 24, 2025. [Online]. Available: https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/
[8]. "Vector Databases for generative AI applications," Community.aws. Accessed: Feb. 24, 2025. [Online]. Available: https://community.aws/content/2f5dkpj96MDM6Y9lumYPjZAB8SX/vector-databases-for-generative-ai-applications
[9]. "10 top vector database options for similarity searches TechTarget," Search Data Management. Accessed: Feb. 24, 2025. [Online]. Available: https://www.techtarget.com/searchdatamanagement/tip/Top-vector-database-options-for-similarity-searches
[10]. "Vector databases in LLMs and search," InfoWorld. Accessed: Feb. 24, 2025. [Online]. Available: https://www.infoworld.com/article/2335281/vector-databases-in-llms-and-search.html
[11]. A. Mittal, "The Role of Vector Databases in Modern Generative AI Applications," Unite.AI. Oct. 2023. Accessed: Feb. 24, 2025. [Online]. Available: https://www.unite.ai/the-role-of-vector-databases-in-modern-generative-ai-applications/
[12]. "Vector Database vs. Knowledge Graph: Making the Right Choice When Implementing RAG," CIO. Accessed: Feb. 24, 2025. [Online]. Available: https://www.cio.com/article/1308631/vector-database-vs-knowledge-graph-making-the-right-choice-when-implementing-rag.html
[13]. "Vector Databases: Intro, Use Cases, Top 5 Vector DBs." Accessed: Feb. 24, 2025. [Online]. Available: https://www.v7labs.com/blog/vector-databases
[14]. D. Brinkmann, "Iveta Lohovska on Gen AI and Vector Search Qdrant - Qdrant." Accessed: Feb. 24, 2025. [Online]. Available: https://qdrant.tech/blog/gen-ai-and-vector-search/
[15]. "IBM watsonx.data's integrated vector database: Unify, prepare, and deliver your data for AI IBM." Apr. 2024. Accessed: Feb. 24, 2025. [Online]. Available: https://www.ibm.com/new/announcements/ibm-watsonx-data-vector-database-ai-ready-data-management
[16]. "Vector database choices in Vertex AI RAG Engine Generative AI," Google Cloud. Accessed: Feb. 24, 2025. [Online]. Available: https://cloud.google.com/vertex-ai/generative-ai/docs/vector-db-choices
[17]. "Building Enterprise-Ready Generative AI Applications with Vertex AI Vector Search." Mar. 2024. Accessed: Feb. 24, 2025. [Online]. Available: https://www.googlecloudcommunity.com/gc/Community-Blogs/Building-Enterprise-Ready-Generative-AI-Applications-with-Vertex/ba-p/723378
[18]. markjbrown, "Integrated vector database - Azure Cosmos DB." Dec. 2024. Accessed: Feb. 24, 2025. [Online]. Available: https://learn.microsoft.com/en-us/azure/cosmos-db/vector-database

DOI: 10.17148/IARJSET.2025.12210

[19]. "The role of vector databases in generative AI applications AWS Database Blog." Jul. 2023. Accessed: Feb. 24, 2025. [Online]. Available: https://aws.amazon.com/blogs/database/the-role-of-vector-datastores-in-generative-ai-applications/

[20]. "Importance of Vector data stores for Gen AI Applications AWS Partner Network (APN) Blog." Nov. 2024. Accessed: Feb. 24, 2025. [Online]. Available: https://aws.amazon.com/blogs/apn/importance-of-vector-data-stores-for-gen-ai-applications/

[21]. "Mosaic AI Vector Search Databricks Documentation." Feb. 2025. Accessed: Feb. 24, 2025. [Online]. Available: https://docs.databricks.com/aws/en/generative-ai/vector-search

[22]. "Snowflake Cortex." Accessed: Feb. 24, 2025. [Online]. Available: https://www.snowflake.com/content/snowflake-site/global/en/product/features/cortex

[23]. "New Vector Database in Salesforce Data Cloud Will Power AI, Analytics, and Automation Using LLMs with Business Data for Use Across the Einstein 1 Platform," Salesforce. Dec. 2023. Accessed: Feb. 24, 2025. [Online]. Available: https://www.salesforce.com/news/press-releases/2023/12/14/unstructured-data-ai-search-einstein/

[24]. "Vector Databases :: Spring AI Reference." Accessed: Feb. 24, 2025. [Online]. Available: https://docs.spring.io/spring-ai/reference/api/vectordbs.html

[25]. sophialagerkranspandey, "AI Integrations for Semantic Kernel." Nov. 2024. Accessed: Feb. 24, 2025. [Online]. Available: https://learn.microsoft.com/en-us/semantic-kernel/concepts/ai-services/integrations

[26]. "What Is A Vector Database? IBM." Jul. 2024. Accessed: Feb. 24, 2025. [Online]. Available: https://www.ibm.com/think/topics/vector-database

[27]. Attri, "Understanding Vector Databases in Generative AI Evolution," Medium. Dec. 2023. Accessed: Feb. 24, 2025. [Online]. Available: https://medium.com/@attriai/understanding-vector-databases-in-generative-ai-evolution-7014190a682a

[28]. B. Ghosh, "Power of Vector Databases for Gen AI Applications," Medium. Sep. 2023. Accessed: Feb. 24, 2025. [Online]. Available: https://medium.com/@bijit211987/power-of-vector-databases-for-gen-ai-applications-a63d4cf7e352

[29]. "Unveiling the Power of Vector Databases in Gen AI Solutions LinkedIn." Accessed: Feb. 24, 2025. [Online]. Available: https://www.linkedin.com/pulse/unveiling-power-vector-databases-gen-ai-solutions-dataenginenz-4yk0c/

[30]. "Beyond GenAI: What Is A Vector Database, And Why Do You Need One? LinkedIn." Accessed: Feb. 24, 2025. [Online]. Available: https://www.linkedin.com/pulse/beyond-genai-what-vector-database-why-do-you-need-one-celia-btppf/

[31]. E. Wallace, "How Vector Databases Enhance GenAI," RTInsights. Nov. 2023. Accessed: Feb. 24, 2025. [Online]. Available: https://www.rtinsights.com/how-vector-databases-enhance-genai/

[32]. "Discover New Gen AI Google Cloud Database Capabilities," Google Cloud Blog. Accessed: Feb. 24, 2025. [Online]. Available: https://cloud.google.com/blog/products/databases/discover-new-gen-ai-google-cloud-database-capabilities

[33]. "Vector Database," Databricks. Oct. 2024. Accessed: Feb. 24, 2025. [Online]. Available: https://www.databricks.com/glossary/vector-database

[34]. "Generative AI," Graph Database & Analytics. Feb. 2025. Accessed: Feb. 24, 2025. [Online]. Available: https://neo4j.com/generativeai/

[35]. A. Srivastava, "Choosing a Vector Database for Your Gen AI Stack SingleStoreDB for AI & Vectors," SingleStore. Jun. 2023. Accessed: Feb. 24, 2025. [Online]. Available: https://www.singlestore.com/blog/choosing-a-vector-database-for-your-gen-ai-stack/

[36]. C. Latimer, "The Ultimate Guide To Vector Database Success In AI - Vectorize." Apr. 2024. Accessed: Feb. 24, 2025. [Online]. Available: https://vectorize.io/what-is-a-vector-database/

[37]. A. T. Williams, "Top 5 Vector Database Solutions for Your AI Project," The New Stack. Jun. 2023. Accessed: Feb. 24, 2025. [Online]. Available: https://thenewstack.io/top-5-vector-database-solutions-for-your-ai-project/

[38]. I. Novogroder, "What is a Vector Database? Top 12 Use Cases," Git for Data - lakeFS. Jul. 2023. Accessed: Feb. 24, 2025. [Online]. Available: https://lakefs.io/blog/what-is-vector-databases/

[39]. "The AI Database Landscape: Vector Search, Gen AI, and More," Aerospike. Accessed: Feb. 24, 2025. [Online]. Available: https://aerospike.com/blog/ai-database-landscape/

[40]. "Generative-ai-for-beginners/15-rag-and-vector-databases/README.md at main · microsoft/generative-ai-for-beginners," GitHub. Accessed: Feb. 24, 2025. [Online]. Available: https://github.com/microsoft/generative-ai-for-beginners/blob/main/15-rag-and-vector-databases/README.md

[41]. "Vector Database Fundamentals," Coursera. Accessed: Feb. 24, 2025. [Online]. Available: https://www.coursera.org/specializations/vector-database-fundamentals

DOI: 10.17148/IARJSET.2025.12210

[42]. Satyadhar Joshi, "A Literature Review of Gen AI Agents in Financial Applications: Models and Implementations," International Journal of Science and Research (IJSR), doi: https://www.doi.org/10.21275/SR25125102816.

[43]. Satyadhar Joshi, "Advancing innovation in financial stability: A comprehensive review of ai agent frameworks, challenges and applications," World Journal of Advanced Engineering Technology and Sciences, vol. 14, no. 2, pp. 117–126, 2025, doi: 10.30574/wjaets.2025.14.2.0071.

[44]. Satyadhar Joshi, "Agentic Generative AI and the Future U.S. Workforce: Advancing Innovation and National Competitiveness," Feb. 03, 2025, Social Science Research Network, Rochester, NY: 5126922. doi: 10.2139/ssrn.5126922.

[45]. Satyadhar Joshi, "Generative AI: Mitigating Workforce and Economic Disruptions While Strategizing Policy Responses for Governments and Companies," Feb. 12, 2025, Social Science Research Network, Rochester, NY: 5135229. doi: 10.2139/ssrn.5135229.

[46]. Satyadhar Joshi, "Implementing Gen AI for Increasing Robustness of US Financial and Regulatory System," IJIREM, vol. 11, no. 6, Art. no. 6, Jan. 2025, doi: 10.55524/ijirem.2024.11.6.19.

[47]. Satyadhar Joshi, "Leveraging prompt engineering to enhance financial market integrity and risk management," World J. Adv. Res. Rev., vol. 25, no. 1, pp. 1775–1785, Jan. 2025, doi: 10.30574/wjarr.2025.25.1.0279.

[48]. Satyadhar Joshi, "Retraining US Workforce in the Age of Agentic Gen AI: Role of Prompt Engineering and Up- Skilling Initiatives," International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), vol. 5, no. 1, 2025.

[49]. Satyadhar Joshi, "Review of autonomous systems and collaborative AI agent frameworks," International Journal of Science and Research Archive, vol. 14, no. 2, pp. 961–972, 2025, doi: 10.30574/ijsra.2025.14.2.0439.

[50]. Satyadhar Joshi, "Review of Data Engineering and Data Lakes for Implementing GenAI in Financial Risk A Comprehensive Review of Current Developments in GenAI Implementations," Jan. 01, 2025, Social Science Research Network, Rochester, NY: 5123081. doi: 10.2139/ssrn.5123081. Doi: https://doi.org/10.48175/IJARSCT-23272

[51]. Satyadhar Joshi, "Review of Data Engineering Frameworks (Trino and Kubernetes) for Implementing Generative AI in Financial Risk," Int. J. Res. Publ. Rev., vol. 6, no. 2, pp. 1461–1470, Feb. 2025, doi: 10.55248/gengpi.6.0225.0756.

[52]. Satyadhar Joshi, "Review of Data Pipelines and Streaming for Generative AI Integration: Challenges, Solutions, and Future Directions", International Journal of Research Publication and Reviews, Vol 6, no 2, pp 2348-2357 February 2025.

[53]. Satyadhar Joshi, "The Synergy of Generative AI and Big Data for Financial Risk: Review of Recent Developments," IJFMR - International Journal For Multidisciplinary Research, vol. 7, no. 1, doi: https://doi.org/g82gmx.

[54]. [49] Satyadhar Joshi, "The Transformative Role of Agentic GenAI in Shaping Workforce Development and Education in the US," Feb. 01, 2025, Social Science Research Network, Rochester, NY: 5133376. Accessed: Feb. 17, 2025. [Online]. Available: https://papers.ssrn.com/abstract=5133376

[55]. Satyadhar Joshi, "Review of Data Engineering Frameworks (Trino and Kubernetes) for Implementing Generative AI in Financial Risk," Int. J. Res. Publ. Rev., vol. 6, no. 2, pp. 1461–1470, Feb. 2025, doi: 10.55248/gengpi.6.0225.0756.

[56]. Satyadhar Joshi, "Review of autonomous systems and collaborative AI agent frameworks," International Journal of Science and Research Archive, vol. 14, no. 2, pp. 961–972, 2025, doi: 10.30574/ijsra.2025.14.2.0439.

