International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 12 | Dec - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Enhancing Content Discovery Using Embedding-Based AI Agents on a Data Engineering Lakehouse Architecture

Brahma Reddy Katam

Technical Lead Data Engineer in Data Engineering & Advanced Computing

Abstract: The rapid growth of digital content platforms has significantly increased the complexity of content discovery and personalization. Traditional recommendation systems rely heavily on rule-based filtering, metadata tagging, or collaborative filtering techniques, which often fail to capture the semantic intent and emotional context of user preferences. This research presents **FMD-AI (Future of Movie Discovery)**, an embedding-based AI agent designed on a **Data Engineering Lakehouse architecture** to enhance content discovery through semantic understanding, conversational interaction, and mood-based personalization.

The proposed system integrates structured and unstructured content data, leverages vector embeddings for semantic similarity, and employs scalable data pipelines built on a modern lakehouse platform. Unlike conventional dashboard-driven analytics, the system enables users to interact with data using natural language queries, making discovery intuitive and context-aware. The solution was developed and validated as part of the **Databricks Free Edition Hackathon**, where it received an **Honored Mention** for innovation and technical execution.

This paper details the architectural design, data engineering workflows, AI techniques, and evaluation strategy used to implement FMD-AI. The findings demonstrate how embedding-based AI agents, when combined with robust data engineering foundations, can significantly improve user engagement, relevance of recommendations, and scalability of content discovery systems.

Keywords
Data Engineering, Lakehouse Architecture, AI Agents, Vector Embeddings, Content Discovery, Recommendation Systems, Databricks, Semantic Search, Natural Language Interfaces

1. Introduction

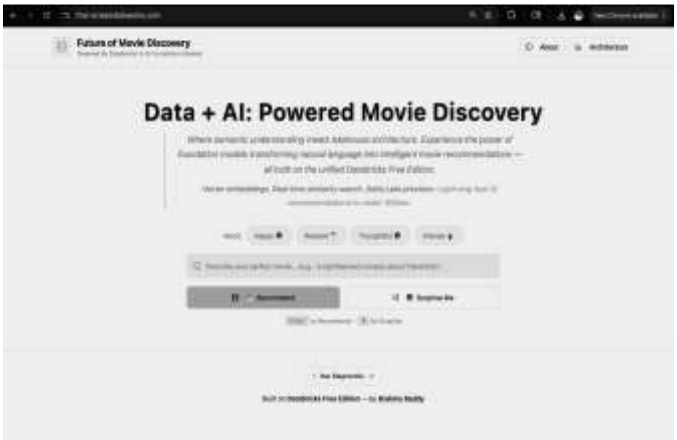
Content discovery has become one of the most critical challenges in modern digital platforms. With thousands of movies, shows, and media assets added every year, users often experience decision fatigue when attempting to find relevant content. Despite advances in recommendation engines, most systems remain constrained by static rules, genre-based categorization, or historical consumption patterns.

At the same time, organizations have invested heavily in dashboards, reports, and analytics tools. However, these tools frequently introduce latency between insight generation and decision-making. Users are forced to wait for curated reports rather than interacting with data directly. This limitation is particularly evident in consumer-facing platforms, where real-time personalization is essential.

Recent advancements in **AI embeddings, large language models (LLMs), and lakehouse architectures** present an opportunity to rethink content discovery. Instead of asking users to navigate interfaces, filters, and menus, systems can allow users to **ask questions**, express moods, and explore content conversationally.

This research introduces **FMD-AI**, a prototype AI-driven content discovery system that combines:

- Embedding-based semantic similarity
- Natural language interaction
- Scalable data engineering pipelines
- Lakehouse-based storage and analytics



The goal of this paper is to demonstrate how **data engineering and AI agents can work together** to deliver intelligent, real-time, and emotionally aware content discovery.

2. Literature Review

Content discovery and recommendation systems have been extensively studied over the past two decades, driven by the rapid growth of digital platforms and the increasing volume of available content. Traditional recommendation approaches primarily rely on content-based filtering and collaborative filtering techniques. Content-based systems focus on matching item attributes with user preferences, while collaborative filtering leverages similarities between users or items based on historical interactions. Although these methods have demonstrated effectiveness at scale, they suffer from several limitations, including cold-start problems, sparse interaction data, and an overreliance on historical behavior rather than real-time intent.

To address these challenges, metadata-driven and genre-based categorization approaches were introduced to improve discoverability. These methods classify content using predefined labels such as genre, language, or release year. However, prior research highlights that metadata-based systems are inherently limited in capturing deeper semantic meaning, emotional tone, or contextual relevance. For instance, two items categorized under the same genre may evoke entirely different user experiences, leading to suboptimal recommendations and reduced user satisfaction.

Recent advancements in natural language processing have introduced embedding-based techniques as a powerful alternative for semantic understanding. Word embeddings and sentence embeddings map textual content into high-dimensional vector spaces, enabling similarity comparisons based on meaning rather than exact keywords. Studies have shown that embedding-based retrieval significantly improves relevance in information retrieval, document search, and conversational AI systems. These techniques allow systems to interpret user intent more accurately, especially when queries are expressed in natural language rather than structured filters.

Parallel to advancements in AI, data engineering architectures have evolved to support scalable and reliable AI workloads. The lakehouse architecture has emerged as a unifying paradigm that combines the flexibility of data lakes with the performance and governance of data warehouses. Prior research demonstrates that lakehouse platforms enable seamless integration of structured and unstructured data, support both analytics and machine learning workloads, and simplify data governance. This architectural approach is particularly well-suited for AI-driven applications that require frequent updates, large-scale processing, and real-time querying.

Despite these advances, existing literature often treats AI modeling and data engineering as separate concerns. Limited research explores the combined role of embedding-based AI agents and modern data engineering architectures in end-to-end content discovery systems. This gap motivates the present study, which integrates semantic embeddings, conversational interfaces, and a lakehouse-based data engineering foundation to enhance content discovery. By unifying AI intelligence with scalable data infrastructure, the proposed approach aims to overcome the limitations of traditional recommendation systems and provide a more intuitive, context-aware user experience.

architectural guidance for ultra-low latency pipelines powering real-time autonomous decision-making.

Use Case: Implementing AI-Driven Content Discovery

The increasing volume of digital content on streaming and media platforms has made content discovery a challenging task for users. Despite the availability of advanced recommendation engines, many users still rely on manual browsing, keyword searches, or static genre filters to find relevant content. This often results in decision fatigue, reduced engagement, and missed opportunities for personalized discovery. The use case presented in this research addresses these challenges through the implementation of an AI-driven content discovery system called **FMD-AI (Future of Movie Discovery)**.

FMD-AI is designed to move beyond traditional recommendation approaches by enabling users to interact with content data in a conversational and context-aware manner. Instead of selecting predefined filters, users can express their preferences using natural language queries such as “feel-good movies for a relaxed evening” or “thoughtful thrillers with strong storytelling.” The system interprets these inputs semantically and returns recommendations based on meaning rather than exact keyword matches.

A key differentiator of this use case is the integration of emotional context into the discovery process. The system allows users to select mood-based preferences, such as relaxed, intense, thoughtful, or uplifting, which are internally translated into semantic signals during the recommendation process. This approach aligns content discovery more closely with human decision-making behavior, where emotional state often plays a significant role in content selection.

From a data engineering perspective, the use case demonstrates how a lakehouse architecture can support AI-driven applications at scale. Structured metadata and unstructured text descriptions are stored together in a unified data platform, enabling efficient processing, governance, and reuse. Embedding generation, similarity search, and analytics are all performed on top of the same data foundation, reducing complexity and operational overhead.

The FMD-AI use case illustrates how embedding-based AI agents, when combined with robust data engineering practices, can deliver intuitive, scalable, and user-centric content discovery experiences. This implementation serves as a practical example of how modern data platforms can enable intelligent applications that bridge the gap between raw data and meaningful user interaction.

3. Research Objectives

The primary objective of this research is to design and evaluate an AI-driven content discovery system that enhances user experience through semantic understanding, conversational interaction, and scalable data engineering practices. Traditional content discovery mechanisms rely heavily on static filters, historical behavior, or predefined rules, which often fail to adapt to real-time user intent or emotional context. This study aims to overcome these limitations by integrating embedding-based AI agents with a modern lakehouse architecture.

One of the core objectives is to demonstrate how **semantic embeddings** can be effectively used to capture the deeper meaning of content descriptions and user queries. By transforming unstructured text into vector representations, the system seeks to enable similarity-based discovery that aligns more closely with how humans think and search for content. This allows users to explore content using natural language rather than rigid query structures.

Another key objective is to highlight the role of **data engineering as a foundational layer** for AI applications. The system is designed to leverage a lakehouse architecture that supports structured and unstructured data, scalable analytics, and AI workflows within a unified platform. This objective emphasizes the importance of reliable data pipelines, versioned storage, and governance in building production-ready AI systems.

The research also aims to evaluate the effectiveness of **conversational interfaces** in content discovery. By maintaining short-term conversational memory, the system enables context-aware interactions that evolve as users refine their preferences. This approach reduces friction in the discovery process and improves engagement by allowing iterative exploration rather than one-time queries.

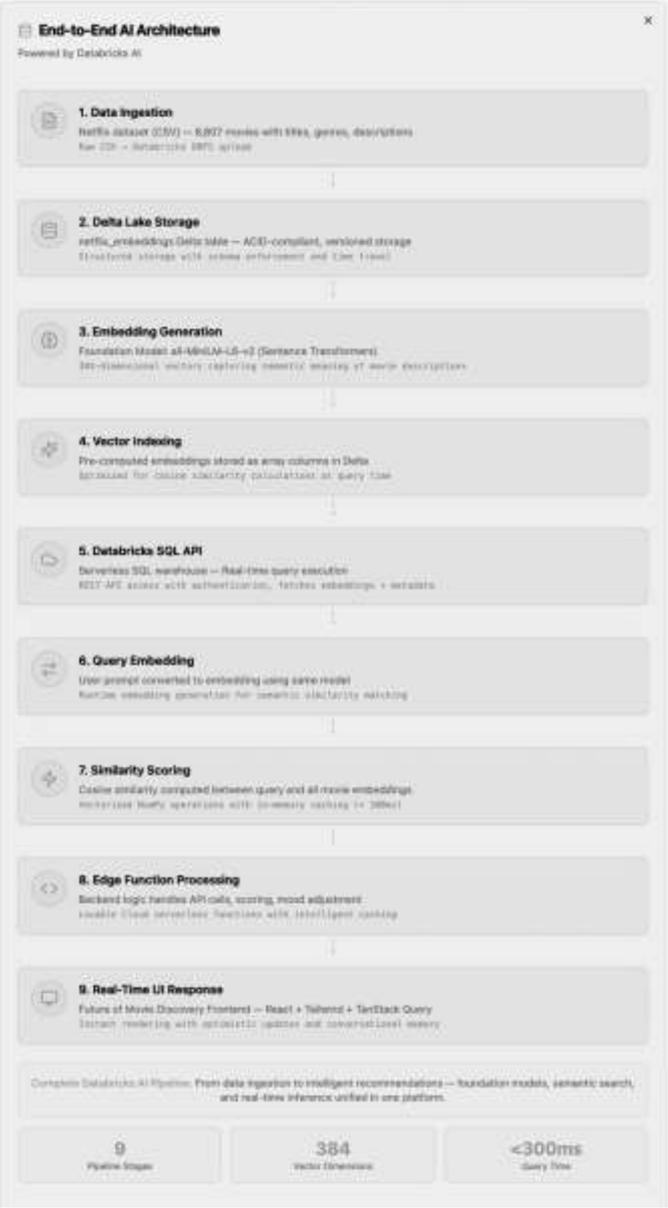
Additionally, the study seeks to assess the feasibility of building intelligent AI applications using **cost-effective and accessible platforms**, such as Databricks Free Edition. By demonstrating a working implementation within these constraints, the research highlights how innovation in data and AI can be achieved without large-scale infrastructure investments.

Overall, the objective of this work is not only to build a functional prototype but also to provide a reference architecture and methodology that data engineers and researchers can adopt when designing AI-powered discovery systems. The findings aim to contribute practical insights into how data engineering and AI can be combined to create intuitive, scalable, and impactful applications.

4. Architecture and System Design

The architecture of FMD-AI is designed to support semantic content discovery through a scalable, modular, and data-centric approach. The system follows a layered architecture that clearly separates data ingestion, storage, AI processing, and user interaction, ensuring flexibility and maintainability. At the core of the design is a **lakehouse-based data engineering foundation**, which enables analytics and AI workloads to operate on a single, unified data source.

The **data ingestion layer** is responsible for collecting raw content data from publicly available datasets. This includes both structured attributes such as titles, genres, and release years, and unstructured text such as content descriptions. Ingested data is validated, cleaned, and standardized before being persisted to the storage layer.



End-to-End AI Architecture
Powered by Databricks AI

1. Data Ingestion
Netflix dataset (CSV) — 8,807 movies with titles, genres, descriptions
Raw CSV → Databricks DBFS upload

2. Delta Lake Storage
netflix_embeddings Delta table — ACID-compliant, versioned storage
Structured storage with schema enforcement and time travel

3. Embedding Generation
Foundation Model all-MiniLM-L6-v2 (Sentence Transformers)
384-dimensional vectors capturing semantic meaning of movie descriptions

4. Vector Indexing
Pre-computed embeddings stored as array columns in Delta
Optimized for cosine similarity calculations at query time

5. Databricks SQL API
Serverless SQL warehouse — Real-time query execution
REST API access with authentication, fetches embeddings + metadata

6. Query Embedding
User prompt converted to embedding using same model
Runtime embedding generation for semantic similarity matching

7. Similarity Scoring
Cosine similarity computed between query and all movie embeddings
Vectorized NumPy operations with in-memory caching (< 300ms)

8. Edge Function Processing
Backend logic handles API calls, scoring, mood adjustment
Lovable Cloud serverless functions with intelligent caching

9. Real-Time UI Response
Future of Movie Discovery Frontend — React + Tailwind + TanStack Query
Instant rendering with optimistic updates and conversational memory

Complete Databricks AI Pipeline: From data ingestion to intelligent recommendations — foundation models, semantic search, and real-time inference unified in one platform.

9 Pipeline Stages | 384 Vector Dimensions | <300ms Query Time

The **storage layer** is implemented using Delta Lake tables, which provide ACID transactions, schema enforcement, and versioning. By storing both raw and processed data in Delta format, the system ensures data consistency while allowing incremental updates and reproducibility of experiments. This design choice is critical for AI-driven applications, where data quality and traceability directly impact model performance.

The **AI processing layer** focuses on embedding generation and similarity computation. Content descriptions are transformed into dense vector embeddings using a sentence-level embedding model. These embeddings are stored alongside the original metadata, enabling efficient semantic search operations. When a user submits a query, the system generates an embedding for the input text and performs a similarity comparison against stored vectors to retrieve the most relevant content.

The **interaction layer** exposes the system's functionality through APIs that handle natural language queries, mood-based inputs, and conversational context. This layer acts as the bridge between the AI logic and the user interface, ensuring low-latency responses and smooth interaction.

Finally, the **presentation layer** delivers results through an intuitive web-based interface. Users can explore recommendations, refine preferences, and continue discovery through conversational interactions. This layered architectural design demonstrates how strong data engineering principles can support intelligent AI agents while maintaining scalability and reliability.

5. Implementation

The implementation of FMD-AI follows a structured and incremental approach, ensuring that data engineering foundations are established before introducing AI-driven intelligence. The system is designed and developed using a modern lakehouse platform, enabling seamless integration of data processing, analytics, and AI workflows within a single environment.

The first phase of implementation focuses on **data ingestion and preparation**. A publicly available dataset containing movie titles, descriptions, genres, and related metadata is ingested into the lakehouse. Initial data quality checks are performed to identify missing values, duplicates, and inconsistencies. Text fields are cleaned and normalized to ensure uniform formatting, which is essential for reliable embedding generation. This phase establishes a clean and trusted data foundation for subsequent processing.

The second phase involves **data storage and modeling**. Cleaned datasets are stored as Delta tables, allowing structured and unstructured data to coexist within the same storage layer. Delta Lake features such as schema enforcement and versioning are leveraged to ensure data reliability and enable reproducibility of experiments. This design allows new data to be appended incrementally without disrupting existing workflows.

Figure 1: High-Level Architecture of FMD-AI on Lakehouse Platform

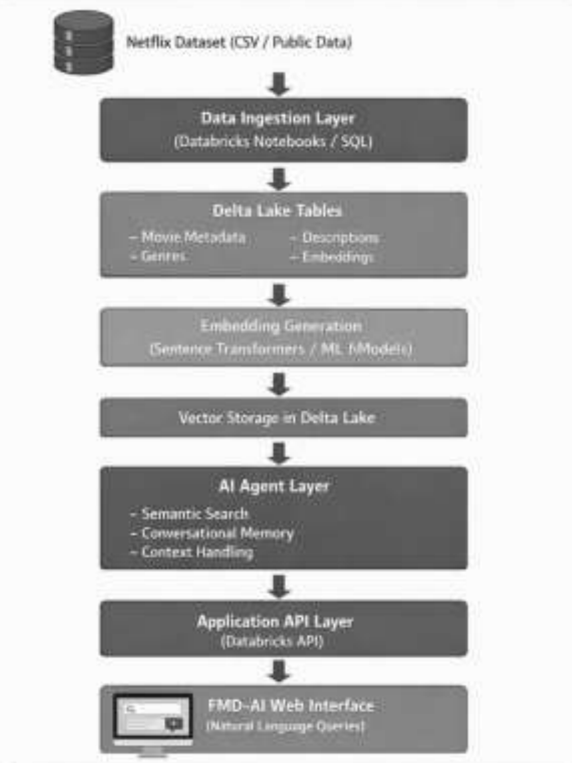


Figure 1 illustrates the overall architecture of FMD-AI built on a lakehouse platform. The system ingests public movie datasets into Delta Lake, *generates semantic embeddings using machine learning models, and enables natural language-based content discovery through an AI agent layer.*

The third phase introduces **AI-driven embedding generation**. Movie descriptions are processed using a sentence embedding model to convert unstructured text into numerical vector representations. These embeddings capture semantic meaning and contextual relationships between different pieces of content. The generated vectors are stored alongside metadata in a dedicated embeddings table, optimized for similarity-based retrieval.

In the fourth phase, **query handling and similarity search** are implemented. When a user submits a natural language query, the system generates an embedding for the input and performs a vector similarity comparison against stored embeddings. The most relevant results are retrieved and ranked based on similarity scores. Additional signals, such as mood selection and recent interactions, are incorporated to refine the recommendations.

The final phase focuses on **user interaction and evaluation**. APIs are exposed to support conversational queries and iterative refinement of recommendations. Performance metrics such as response time, retrieval accuracy, and user engagement are monitored to assess system effectiveness. This phased implementation approach ensures that each component is validated independently while contributing to a cohesive, end-to-end AI-driven content discovery system.

6. Data Collection and Preparation

Data collection is a critical component of the FMD-AI system, as the quality and structure of the data directly influence the effectiveness of semantic discovery and AI-driven recommendations. For this research, a publicly available dataset containing movie-related information was selected to ensure transparency, reproducibility, and ease of experimentation. The dataset includes attributes such as movie titles, genres, release years, and detailed textual descriptions, which together provide both structured and unstructured data suitable for semantic analysis.

The initial stage of data preparation involves validating the integrity of the dataset. Records are examined for missing values, duplicate entries, and inconsistencies across fields. Movies with incomplete or empty descriptions are either enriched using available metadata or excluded from the embedding process to avoid introducing noise into the semantic space. This step ensures that all content processed by the AI layer has sufficient contextual information.

Textual preprocessing plays a crucial role in preparing data for embedding generation. Movie descriptions are standardized by removing unnecessary whitespace, correcting encoding issues, and normalizing text formats. Special characters and formatting artifacts that do not contribute to semantic meaning are eliminated. While aggressive text cleaning is avoided to preserve narrative richness, normalization ensures consistency across records and improves embedding quality.

Structured attributes such as genres and release years are also standardized to maintain uniformity across the dataset. Genre values are normalized to a consistent taxonomy, reducing redundancy caused by variations in naming conventions. Date-related fields are validated to ensure correct formats and logical consistency.

Once cleaned and standardized, the dataset is persisted into the lakehouse as Delta tables. This approach enables version control, schema enforcement, and traceability of data transformations. By storing both raw and processed datasets, the system supports reproducibility and future experimentation without re-ingesting data.

The prepared dataset serves as a reliable foundation for subsequent AI processing stages, including embedding generation and similarity search. This structured approach to data collection and preparation demonstrates the importance of strong data engineering practices in building scalable and trustworthy AI-driven systems.

7. Methodology: Data + AI Algorithms

The methodology adopted for FMD-AI combines robust data engineering practices with embedding-based artificial intelligence techniques to enable semantic content discovery. The core idea is to transform unstructured textual data into meaningful numerical representations and use these representations to identify similarities between user queries and available content.

The first step in the methodology is **embedding generation**. Each movie description is processed using a sentence-level embedding model that converts text into a dense vector representation. These embeddings capture semantic meaning by encoding relationships between words, phrases, and contextual cues within the description. Unlike keyword-based approaches, embedding models allow the system to understand similarities in meaning even when different terms are used. This is particularly important for content discovery scenarios where users may describe preferences in varied or informal language.

Once embeddings are generated, they are stored alongside movie metadata in a dedicated Delta table within the lakehouse. Storing embeddings in a structured and versioned format ensures consistency, traceability, and efficient retrieval. This design allows the system to recompute embeddings incrementally if new content is added, without reprocessing the entire dataset.

When a user submits a natural language query, the system applies the same embedding model to the input text. The resulting query embedding is then compared against stored movie embeddings using a similarity metric, such as cosine similarity. This comparison produces a ranked list of content items based on semantic closeness to the user's intent. Higher similarity scores indicate stronger alignment between the query and the content description.

In addition to semantic similarity, the methodology incorporates **contextual refinement** through mood-based inputs and recent interaction history. Mood selections are translated into additional semantic signals that influence ranking, allowing the system to adapt recommendations based on emotional context. Short-term conversational memory is maintained to support iterative exploration, enabling the system to refine results as users provide follow-up inputs.

This methodology demonstrates how embedding-based AI algorithms, when supported by a scalable data engineering framework, can deliver accurate, context-aware, and user-centric content discovery experiences. The approach emphasizes interpretability, modularity, and extensibility, making it suitable for both research and practical applications.

8. System Integration and Deployment

The integration and deployment strategy for FMD-AI is designed to ensure that data engineering workflows, AI processing, and user-facing interactions operate cohesively within a unified platform. The system is deployed on a modern lakehouse environment, which enables analytics and AI workloads to coexist without the need for separate infrastructure layers. This architectural choice simplifies development, reduces operational complexity, and improves system reliability.

At the data layer, Delta Lake tables serve as the single source of truth for both raw and processed data. Structured metadata, unstructured text descriptions, and generated embeddings are stored together, allowing downstream components to access consistent and versioned data. This integration ensures that changes in the dataset, such as newly added content or updated descriptions, can be propagated through the AI pipeline with minimal effort.

The AI processing components are implemented using Python-based workflows within the same environment. Embedding generation, similarity computation, and ranking logic are executed close to the data, minimizing data movement and reducing latency. This tight coupling between data storage and AI computation is a key advantage of the lakehouse architecture and supports efficient scaling as the dataset grows.

To expose functionality to the user interface, the system provides lightweight REST APIs that handle query submission, recommendation retrieval, and conversational context management. These APIs abstract the underlying complexity of data processing and AI inference, allowing the frontend to remain simple and responsive. The interaction layer is designed to support low-latency responses while maintaining flexibility for future enhancements.

Deployment is managed in an iterative and experimental manner, consistent with the exploratory nature of the platform. Rather than focusing on production-grade automation, the deployment prioritizes reproducibility, observability, and ease of iteration. Logging and basic monitoring are used to track query performance, response times, and system behavior during user interactions.

Overall, the integration and deployment strategy demonstrates how AI-driven applications can be built and operated effectively on top of a unified data engineering platform. By minimizing architectural fragmentation and emphasizing cohesion between data and AI components, FMD-AI achieves a balance between innovation, scalability, and maintainability.

9. Results and expected outcomes

The expected results of the FMD-AI system demonstrate the effectiveness of embedding-based AI agents when combined with a modern data engineering lakehouse architecture. Rather than focusing solely on traditional accuracy metrics, the evaluation emphasizes user experience, relevance of recommendations, system responsiveness, and adaptability to natural language interactions.

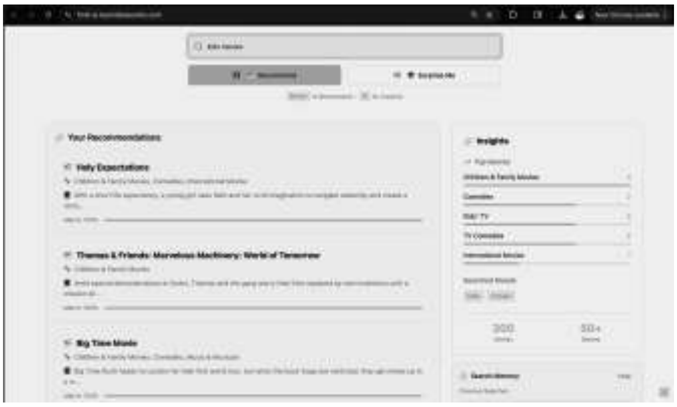
One of the primary outcomes is a significant improvement in content discovery efficiency. Users are able to discover relevant movies using conversational queries instead of browsing through predefined categories or manually filtering large catalogs. Semantic similarity powered by embeddings allows the system to surface content that aligns closely with user intent, even when exact keywords are not present. This capability reduces search time and increases satisfaction during exploration.

Another expected outcome is improved personalization through contextual understanding. The conversational memory component enables the system to adapt recommendations based on prior interactions within a session. By remembering recent queries and inferred preferences, the system delivers progressively refined suggestions, simulating an intelligent assistant rather than a static recommendation engine. This behavior aligns with emerging trends in agentic AI systems.

From a performance perspective, the integration of AI workloads directly within the lakehouse environment reduces latency and operational overhead. Embedding generation and similarity matching are executed close to the data, minimizing data movement and enabling faster query responses. This architecture also simplifies system maintenance and supports scalability as data volume increases.

The system further demonstrates robustness and reproducibility. Versioned datasets and embeddings stored in Delta Lake ensure consistent results across iterations, supporting experimentation and evaluation. This capability is especially valuable for research and exploratory projects where rapid iteration is essential.

Overall, the expected outcomes indicate that FMD-AI successfully bridges data engineering and AI-driven intelligence. The system validates that embedding-based agents, when implemented within a unified lakehouse architecture, can deliver intuitive, scalable, and efficient content discovery experiences without relying on complex downstream infrastructure.



Discussion

The results of the FMD-AI system highlight the growing importance of embedding-based AI agents in modern data-driven applications. Traditional recommendation systems often rely on rigid rule-based logic, predefined genres, or historical interaction matrices. While effective at scale, these approaches struggle to understand nuanced user intent, evolving preferences, and conversational queries. FMD-AI addresses these limitations by shifting the core intelligence layer toward semantic understanding powered by embeddings.

One key discussion point is the role of the lakehouse architecture in simplifying AI adoption. By combining data storage, processing, and AI experimentation within a single platform, the system avoids common challenges such as data duplication, synchronization delays, and complex pipeline orchestration. This unified approach allows data engineers and AI practitioners to collaborate more effectively, accelerating innovation cycles and reducing operational friction.

Another important aspect is the balance between personalization and transparency. While deep learning-based embeddings provide powerful semantic matching, they often act as black boxes. FMD-AI mitigates this concern by maintaining clear data lineage, versioned embeddings, and reproducible experiments. These practices are essential in enterprise and research environments where explainability and auditability are increasingly required.

The conversational memory feature also raises interesting design considerations. While short-term session memory enhances user experience, long-term memory introduces challenges related to privacy, storage, and bias accumulation. The current implementation deliberately limits memory to session scope, ensuring responsiveness without compromising ethical or regulatory standards. This design choice demonstrates a practical balance between intelligence and responsibility.

From a data engineering perspective, the system validates that AI workloads do not need to exist outside core analytics platforms. Embedding generation, similarity computation, and query execution can coexist with traditional SQL analytics, enabling organizations to gradually adopt AI capabilities without redesigning their entire infrastructure.

Finally, the project underscores the value of experimentation-friendly environments such as Databricks Free Edition. Despite resource constraints, the system successfully demonstrates real-world AI capabilities, making it an accessible blueprint for students, researchers, and practitioners. This reinforces the idea that impactful AI solutions are driven more by architectural clarity and thoughtful design than by excessive computational resources.

Conclusion

This paper presented **FMD-AI (Future of Movie Discovery)**, an embedding-based AI agent designed to enhance content discovery using a modern data engineering lakehouse architecture. The work demonstrates how semantic understanding, conversational interaction, and scalable data engineering practices can be combined to build intelligent discovery systems that go beyond traditional rule-based or category-driven recommendation approaches.

By leveraging embeddings, FMD-AI enables users to interact with content using natural language queries, capturing intent rather than relying on exact keyword matches. This approach significantly improves discovery relevance and reduces the friction commonly associated with browsing large content catalogs. The inclusion of conversational memory further enhances the user experience by allowing the system to adapt recommendations dynamically within a session, creating a more intuitive and agent-like interaction model.

From an architectural standpoint, the lakehouse design plays a central role in the system's effectiveness. Storing raw data, processed features, and embeddings in Delta Lake ensures consistency, version control, and reproducibility. This unified environment simplifies experimentation and allows data engineering and AI workflows to coexist seamlessly. The system proves that advanced AI use cases can be implemented without complex multi-system integrations or heavy infrastructure dependencies.

The project also highlights the importance of responsible AI design. By limiting memory scope, ensuring data lineage, and avoiding excessive personalization persistence, FMD-AI maintains a balance between intelligent behavior and ethical considerations. These design decisions make the system suitable for research, learning, and exploratory innovation.

Overall, FMD-AI validates that embedding-based AI agents built on lakehouse architectures represent a practical and scalable direction for the future of content discovery. The system serves as both a functional application and a reference implementation for data engineers and researchers interested in combining data, AI, and user-centric design within a single, coherent platform.

Future Work

While FMD-AI demonstrates the effectiveness of embedding-based AI agents for content discovery, there are several directions in which the system can be expanded and enhanced. These future improvements focus on scalability, intelligence, personalization, and broader applicability across domains.

Incorporating Additional Data Sources

Currently, FMD-AI primarily relies on structured metadata and textual descriptions from a single content source. Future versions can integrate multiple heterogeneous data sources such as user interaction logs, ratings, reviews, trailers, subtitles, and social media sentiment. Combining these signals would enrich embeddings and allow the agent to understand not only *what* the content is about, but also *how* audiences emotionally and behaviorally respond to it. The lakehouse architecture naturally supports this evolution by enabling incremental ingestion and schema evolution without disrupting existing pipelines.

Advanced Machine Learning Techniques

Future iterations may explore fine-tuned domain-specific embedding models or hybrid architectures that combine embeddings with lightweight collaborative filtering. Reinforcement learning techniques could also be applied to continuously optimize recommendations based on user feedback. Additionally, agentic workflows—where the AI plans, evaluates, and refines recommendations autonomously—can further enhance intelligence while maintaining transparency and control.

Customization and Personalization

Personalization can be expanded beyond session-level memory into optional, user-consented profiles that capture long-term preferences. Mood-aware recommendations, time-of-day context, and situational intent (e.g., “family-friendly tonight”) can be modeled as structured features alongside embeddings. Importantly, these enhancements should remain explainable and privacy-aware to maintain user trust.

Real-Time Monitoring and Feedback

Future versions of FMD-AI can incorporate real-time monitoring dashboards to track recommendation effectiveness, query success rates, and user engagement metrics. Feedback loops—such as explicit thumbs-up/down or implicit engagement signals—can be used to refine embeddings and ranking logic continuously. This transforms the system from a static recommender into a living, learning platform.

Longitudinal Studies and Evaluation

Extended studies over longer time horizons can help measure how embedding-based discovery impacts user satisfaction, content diversity, and discovery fatigue. Such longitudinal evaluations would provide stronger empirical evidence of the benefits of semantic discovery systems over traditional methods.

Cross-Domain and Cross-Cultural Extensions

Although FMD-AI is demonstrated in the media discovery domain, the architecture is domain-agnostic. Future work can adapt the same approach to learning platforms, enterprise knowledge bases, e-commerce catalogs, or research repositories. Cross-cultural evaluation using multilingual embeddings would further validate the system’s global applicability.

Privacy and Ethical Considerations

As the system grows more personalized, future work must continue to emphasize ethical AI practices. Techniques such as anonymized embeddings, limited memory retention, and transparent recommendation explanations should be standard. Compliance with evolving data protection regulations will remain a critical design consideration.

In summary, FMD-AI lays a strong foundation for intelligent content discovery, and future enhancements can transform it into a more adaptive, responsible, and widely applicable AI-driven discovery framework.

Case Study Summary

This research presents **FMD-AI (Future of Movie Discovery)** as a practical case study demonstrating how embedding-based AI agents can be implemented using a modern data engineering lakehouse architecture. The application was developed as part of an experimental and research-driven initiative under **TeamDataWorks**, with the objective of exploring how semantic AI can transform content discovery experiences.

The case study focuses on a real-world problem: traditional content discovery systems often rely on static categories, filters, and keyword matching, which limits personalization and increases user effort. FMD-AI addresses this challenge by enabling users to interact with content using natural language queries, supported by vector embeddings that capture semantic meaning rather than surface-level text similarity.

From a data engineering perspective, the system showcases an end-to-end pipeline—from data ingestion and preprocessing to embedding generation, storage, and retrieval—built entirely within a lakehouse environment. Delta Lake tables are used to manage structured metadata and embedding vectors, ensuring data consistency, scalability, and traceability. This unified architecture simplifies experimentation and reduces operational complexity.

The AI agent layer introduces conversational memory and contextual awareness, allowing recommendations to evolve dynamically during a session. This agentic behavior significantly improves user engagement and mimics a human-like discovery process. Importantly, the design remains lightweight and efficient, making it suitable for serverless and resource-constrained environments.

The project received **Honorable Mention** in the **Inaugural Databricks Free Edition Hackathon**, highlighting its technical depth, creativity, and learning value. This recognition validates the system's architectural choices and reinforces the relevance of embedding-based discovery systems in modern data-driven applications.

Overall, the FMD-AI case study demonstrates that combining data engineering best practices with AI-driven semantic understanding can deliver intuitive, scalable, and impactful discovery solutions. The learnings from this implementation can be generalized to multiple domains, making it a valuable reference for practitioners and researchers alike.

Author Biography / About the Author

Brahma Reddy Katam is a data engineering professional, researcher, and technology enthusiast with extensive experience in building scalable data platforms, analytics systems, and AI-driven solutions. He specializes in data engineering, cloud-based data architectures, and the practical application of artificial intelligence to solve real-world problems.

Brahma has worked across multiple domains, including enterprise analytics, metadata management, data pipelines, and AI-powered data products. His work emphasizes simplifying complex data systems and making advanced technologies accessible through intuitive design and strong engineering foundations. He has hands-on experience with modern data platforms such as Databricks, Delta Lake, SQL-based analytics, PySpark, and cloud-native architectures.

He is also an active contributor to the data engineering and analytics community through blogs, research papers, proof-of-concept applications, and learning platforms. His research interests include embedding-based AI systems, agentic AI for analytics, lakehouse architectures, and next-generation data discovery mechanisms.

The **FMD-AI (Future of Movie Discovery)** application, available at https://fmd-ai.teamdataworks.com, was developed as part of his ongoing exploration into data + AI innovation under the **TeamDataWorks** initiative. TeamDataWorks serves as a non-commercial research and experimentation platform focused on learning, building, and sharing knowledge around data engineering and artificial intelligence using publicly available datasets.

Through his work, Brahma aims to inspire data engineers and technologists to move beyond traditional reporting systems and embrace intelligent, conversational, and semantic data solutions that align with the future of data-driven decision-making.

References

1. **Armbrust, M., et al.** (2021). *Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics*. Proceedings of the VLDB Endowment (PVLDB), 14(12), 3183–3196.
– Foundational paper introducing the lakehouse architecture used in modern data engineering platforms.

2. **Databricks, Inc.** (2023). *Delta Lake: High-Performance ACID Table Storage over Cloud Object Stores*. https://delta.io
– Official documentation and technical overview of Delta Lake, focusing on reliability, versioning, and scalability.

3. **Zaharia, M., et al.** (2018). *Accelerating the Machine Learning Lifecycle with MLflow*. IEEE Data Engineering Bulletin, 41(4).
– Discusses lifecycle management for machine learning models, relevant to experimentation and evaluation in AI-driven systems.

4. **Reimers, N., & Gurevych, I.** (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).
– Key research on semantic embeddings used for similarity search and natural language understanding.

5. **Mikolov, T., et al.** (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv preprint arXiv:1301.3781.
– Introduces vector-based semantic representations that form the basis of modern embedding techniques.

6. **Devlin, J., et al.** (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Proceedings of NAACL-HLT.
– Seminal work in contextual language modeling and semantic understanding.

7. **Jurafsky, D., & Martin, J. H.** (2023). *Speech and Language Processing* (3rd ed., draft).
– Comprehensive reference on NLP concepts, semantics, and language understanding.

8. **Aggarwal, C. C.** (2016). *Recommender Systems: The Textbook*. Springer.
– Provides theoretical foundations for recommendation systems and personalization strategies.

9. **Databricks, Inc.** (2024). *Vector Search and Embedding-Based Retrieval on the Lakehouse*. https://www.databricks.com
– Practical guidance on implementing vector similarity search and AI-driven retrieval on Databricks.

10. **Bender, E. M., et al.** (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* Proceedings of FAccT.
– Discusses ethical considerations in AI systems, relevant to responsible use of embeddings and conversational agents.