International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 04 April-2026 | Impact Factor: 3.5

AI Powered Collaborative LakeHouse Analytics Platform

N. Mani Prem Gowtham
UG Student, Dept of CSE (Data Science)
Vidya Jyothi Institute of Technology
Hyderabad, Telangana, India
nmanipremgowtham@gmail.com

P. Umesh Rayal
UG Student, Dept of CSE (Data Science)
Vidya Jyothi Institute of Technology
Hyderabad, Telangana, India
pogakulaumeshrayal@gmail.com

K. Venkata Pavan Kumar
UG Student, Dept of CSE (Data Science)
Vidya Jyothi Institute of Technology
Hyderabad, Telangana, India
pavankumarkatta26@gmail.com

R. Sai Bharath
UG Student, Dept of CSE (Data Science)
Vidya Jyothi Institute of Technology
Hyderabad, Telangana, India
rallabandisaibharath@gmail.com

Mrs. M. Prasanna Kumari
Assistant Professor,
Dept of CSE (Data Science)
Vidya Jyothi Institute of Technology
Hyderabad, Telangana, India
prasannamcse@vjit.ac.in

ABSTRACT— An Intelligent Lakehouse-Based Data Analytics and Query System is an intelligent data analytics platform designed to simplify the process of extracting meaningful insights from large and complex datasets by integrating Artificial Intelligence with modern data architecture principles. Traditional data analysis systems require users to possess strong technical knowledge and write complex SQL queries, which creates a significant barrier for non-technical users and slows down decision-making. This project addresses these challenges by enabling users to input queries in natural language, which are automatically converted into optimized SQL queries using an AI-based query engine. The system is built on a Lakehouse architecture, which combines the scalability and flexibility of data lakes with the performance and structure of data warehouses, allowing efficient storage and processing of structured, semi-structured, and unstructured data. It incorporates multiple processing engines such as DuckDB for fast SQL-based analytics, Polars for high-performance data processing, and PySpark for handling large-scale distributed datasets, ensuring optimal performance based on the size and complexity of the data. Additionally, the platform includes a smart query validation mechanism to correct errors, a suggestion system to guide users in query formulation, and an intelligent query planner that selects the most suitable execution engine. The results are presented through interactive dashboards featuring KPI visualizations, charts, and dynamic filters, making data interpretation intuitive and accessible. By combining AI-driven query processing, multi-engine execution, and user-friendly visualization, Insight Lake AI reduces the complexity of data analytics, improves efficiency, and enhances decision-making capabilities. This project demonstrates how intelligent systems can effectively bridge the gap between raw data and actionable insights, making advanced analytics accessible, faster, and more efficient for a wide range of users.

INTRODUCTION

In the modern digital era, data is being generated at an unprecedented rate from a wide variety of sources such as mobile applications, websites, cloud services, IoT sensors, social media networks, online transactions, and enterprise information systems. Every click, purchase, interaction, and activity creates valuable data that can be analysed to gain meaningful insights. Organizations use this data to understand customer preferences, improve operational efficiency, monitor performance, forecast future trends, and make strategic decisions. As businesses continue to grow, the volume, variety, and velocity of data also increase, making traditional methods of data management and analytics less effective. Handling such massive datasets with conventional tools often becomes slow, expensive, and difficult to scale. Although data is highly valuable, extracting useful knowledge from it remains a major challenge for many organizations. Traditional analytics platforms generally require users to write complex SQL queries, understand database schemas, and have technical expertise in programming or data engineering tools. This creates a barrier for non-technical users such as managers, executives, domain experts, and business analysts who need quick access to information for decision-making. In many cases, they must depend on IT teams or data specialists to generate reports, which leads to delays, reduced productivity, and slower business responses. Furthermore, many organizations struggle with maintaining separate systems for storage, analytics, and reporting, resulting in higher costs, duplicated efforts, and data inconsistency. To overcome these challenges, Insight Lake AI is proposed as an intelligent, scalable, and user-friendly analytics platform that transforms the way people interact with data. The system combines the power of Artificial Intelligence with modern Lakehouse architecture, which integrates the flexibility of data lakes with the performance and reliability of data warehouses. This unified architecture enables organizations to store, process, and analyse structured, semi-structured, and unstructured data within a single platform. As a result, users no longer need multiple disconnected tools for managing their analytics workflow. One of the key innovations of Insight Lake AI is its natural language query interface. Instead of writing SQL code manually, users can simply ask questions in plain English, such as “What are the top-selling products this month?”, “Show revenue by region,” or “Which customers generated the highest profit?” The AI engine uses Natural Language Processing (NLP) and intelligent query generation techniques to understand user intent, convert the request into SQL automatically, and execute it on the selected dataset. 4 This removes the technical complexity of traditional systems and makes advanced analytics accessible to everyone, regardless of their technical background. It also reduces errors, saves time, and allows users to focus on decision-

making rather than coding. The platform is further strengthened by the integration of multiple high-performance processing engines such as DuckDB, Polars, and PySpark. Each engine is selected based on the workload requirements. DuckDB provides fast SQL execution for local and medium-sized analytical tasks, Polars offers efficient DataFrame operations and transformations, and PySpark enables distributed processing for large-scale big data workloads. This multi-engine design ensures flexibility, speed, and scalability, allowing the system to handle everything from small datasets to enterprise-level data volumes efficiently. Another important feature of Insight Lake AI is its ability to present results through interactive dashboards, visual reports, KPI cards, charts, and graphs. Instead of displaying raw tables or text-based outputs, the system converts query results into clear visualizations that help users quickly understand patterns, trends, and anomalies. Decision-makers can monitor key metrics such as sales growth, profit margins, customer retention, and operational performance in real time. These visual insights improve communication, support faster decision-making, and help organizations respond proactively to market changes. In conclusion, Insight Lake AI represents a next-generation approach to data analytics by combining AIdriven intelligence, natural language interaction, Lakehouse architecture, scalable processing engines, and advanced visualization tools into one unified platform. It eliminates the complexity of traditional analytics systems, empowers both technical and non-technical users, and enables organizations to make faster, smarter, and more data-driven decisions. As the importance of data continues to grow, platforms like Insight Lake AI will play a vital role in shaping the future of intelligent business analytics.

I. PROBLEM DEFINITION

The rapid maturation of large language models (LLMs) presents a compelling opportunity to reimagine the travel booking workflow. An AI-powered chat interface can understand natural language queries such as 'Find me a direct flight from London to New York next Friday, economy class for two adults,' extract all required structured fields, execute the appropriate backend operations, and return a formatted, human-readable response — all within a single, fluid conversational turn.

However, naively delegating business logic to an LLM introduces significant risks: LLMs are prone to hallucination, non-determinism, and unreliable structured output. A robust system must therefore use the LLM narrowly and precisely — for language understanding — while keeping all execution logic firmly in deterministic backend code.

Conventional flight booking platforms — such as those built on legacy GDS (Global Distribution System) interfaces — present users with complex, multi-step form-based workflows. Users must independently navigate origin and destination fields, date selectors, passenger type dropdowns, fare class filters, and seat maps. This paradigm, while functional, creates significant friction, particularly for non-technical users, travellers with complex itineraries, or those unfamiliar with airline jargon such as IATA codes, fare classes, or ticketing rules.

The cognitive load imposed by such interfaces often leads to booking abandonment, user errors, and a poor overall customer experience. Furthermore, traditional systems rarely maintain conversational context — if a user changes their mind mid-flow, they must restart the process entirely rather than simply expressing a new intent.

1.2 PROJECT FEATURES

The chat interface is the primary user touchpoint of the system. Built in React, it renders a real-time message stream resembling popular messaging applications such as WhatsApp or Slack. Users type natural language queries — for example, 'I want to fly from Dubai to London on the 15th of next month, two adults, business class' — and receive structured, formatted responses from the AI agent. The interface maintains a visible message history, displays loading indicators during processing, and supports multi-turn dialogue, allowing users to refine their requests, ask follow-up questions, and navigate the complete booking flow without leaving the chat window.

Related Work

Recent advancements in data engineering and artificial intelligence have led to the emergence of the lakehouse architecture, which integrates the scalability of data lakes with the structured querying capabilities of data warehouses. In [1], the authors introduced the foundational concept of lakehouse systems, highlighting their ability to support both analytics and transactional workloads. However, the proposed system lacks built-in support for collaborative analytics and intelligent automation. In [2], a comprehensive survey of lakehouse architectures was presented, focusing on data ingestion, storage, and query processing mechanisms. While the study provides detailed architectural insights, it does not incorporate advanced AI-driven analytics or real-time collaboration features. Similarly, in [3], the integration of artificial intelligence and machine learning into lakehouse systems was explored to enable predictive analytics and intelligent decision-making. However, the approach lacks interactive and collaborative analytics capabilities. Furthermore, in [4], an AI/ML-optimized lakehouse architecture was proposed to enhance data processing efficiency. Although the system improves performance, it introduces implementation complexity and does not address collaborative workflows. In [5], the evolution of lakehouse systems was discussed, emphasizing features such as ACID transactions, schema enforcement, and governance. Despite these advancements, challenges remain in achieving real-time scalability. Cloud-based implementations were explored in [6], demonstrating the ability of lakehouse systems to handle large-scale distributed data. However, such systems heavily depend on cloud infrastructure and lack integrated collaborative tools. Industry-based approaches discussed in [7] and [16] highlighted cost reduction and unified analytics benefits of lakehouse systems, but these solutions primarily focus on enterprise applications and overlook academic challenges such as explainability and user collaboration. The architectural design and unified data management capabilities of lakehouse systems were further discussed in [8] and [9], emphasizing schema enforcement and real-time analytics. However, these approaches do not incorporate intelligent query handling using AI agents. In [10], a domain-specific implementation in healthcare demonstrated the effectiveness

of lakehouse systems for AI-based analytics, but its applicability is limited and lacks general-purpose collaborative features

II. METHODOLOGY

1. Frontend Layer
The Frontend Layer is the part of the system that users interact with directly. It is developed using Next.js to create a modern, responsive, and user-friendly web application. This layer acts as the communication bridge between users and the analytics system.

2. Backend Layer
The Backend Layer is the core processing unit of the platform. It is developed using Python and FastAPI. This layer receives requests from the frontend, processes them, and returns results. It manages communication between all other layers of the system.

3. AI Query Engine
The AI Query Engine is one of the most important components of Insight Lake AI. It removes the need for users to write SQL queries manually. Instead, users can ask questions in plain English, and the AI converts them into executable SQL statements. This engine uses advanced AI models such as OpenAI or Google Gemini or other AI models.

4. Backend Processing
The Execution Layer is responsible for running queries and processing data. Instead of depending on only one engine, the system supports multiple high-performance engines. Based on the dataset size and workload, the best engine is selected automatically.

5. Database Management:
The Data Storage Layer stores all data used by the platform. It follows the Lakehouse model, which combines the benefits of Data Lakes and Data Warehouses. This allows the system to store structured, semi-structured, and unstructured data in one place.

III. PROPOSED SYSTEM

Recent advancements in data engineering and artificial intelligence have led to the emergence of the lakehouse architecture, which integrates the scalability of data lakes with the structured querying capabilities of data warehouses. In [1], the authors introduced the foundational concept of lakehouse systems, highlighting their ability to support both analytics and transactional workloads. However, the proposed system lacks built-in support for collaborative analytics and intelligent automation. In [2], a comprehensive survey of lakehouse architectures was presented, focusing on data ingestion, storage, and query processing mechanisms. While the study provides detailed architectural insights, it does not incorporate advanced AI-driven analytics or real-time collaboration features. Similarly, in [3], the integration of artificial intelligence and machine learning into lakehouse systems was explored to enable predictive analytics and intelligent decision-making. However, the approach lacks interactive and collaborative analytics capabilities.

IV. IMPLEMENTATION DETAILS

The workflow of Insight Lake AI explains how the system processes a user's request from the moment the query is entered until the final insights are displayed on the dashboard. The platform is designed to make data analysis simple, fast, and intelligent by combining Artificial Intelligence, backend services, multiple data processing engines, and visualization tools. Each step in the workflow has an important role in converting raw user questions into meaningful business insights. This structured workflow ensures accurate query handling, efficient processing, and a smooth user experience. The system allows users to ask questions in natural language instead of writing technical SQL queries. After receiving the request, the platform automatically converts the query into SQL, checks for errors, chooses the best processing engine, executes the query on stored data, and presents the results in the form of tables, charts, dashboards, and KPI cards. This reduces manual effort, saves time, and helps users make better decisions quickly.

4.1 ALGORITHMS USED

The proposed AI-Powered Collaborative Lakehouse Analytics Platform utilizes a combination of intelligent algorithms to enable efficient data processing, query execution, and visualization. The core algorithm is the Natural Language Processing (NLP)-based query generation algorithm, which converts user input in plain English into structured SQL queries. This process involves text preprocessing, intent recognition, and prompt-based query generation using AI models such as OpenAI or Gemini. The system then applies a query validation and optimization algorithm, which checks for syntax errors, verifies schema consistency, and applies optimization techniques to improve query performance and reduce execution time.

To enhance efficiency, the platform employs a dynamic engine selection algorithm, which analyzes the query complexity and dataset size to select the most appropriate execution engine. For small datasets, DuckDB is used due to its fast local analytical capabilities; for medium datasets, Polars is selected for efficient data transformations; and for large datasets, PySpark is utilized for distributed processing. Additionally, a query execution algorithm is implemented to process the validated SQL query on the selected engine and retrieve the results from the Lakehouse storage layer.

After execution, a result processing algorithm formats the output into structured data, calculates key performance indicators (KPIs), and prepares it for visualization. Finally, a visualization mapping algorithm determines suitable chart types and converts the processed data into interactive dashboards using visualization libraries. Together, these algorithms enable an end-to-end intelligent workflow that transforms raw user queries into meaningful insights, improving performance, scalability, and user accessibility.

V. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed AI-Powered Collaborative Lakehouse Analytics Platform was evaluated using multiple datasets of varying sizes and formats, including CSV and Parquet files, to analyze its performance, scalability, and usability. The experimentation focused on assessing the effectiveness of natural language query processing, execution efficiency, and visualization capabilities. The system was tested with different types of user queries, ranging from simple aggregations to complex analytical queries, to ensure robustness and accuracy. The AI-based query engine successfully converted natural language inputs into valid SQL queries with high accuracy, significantly reducing the need for manual intervention and technical expertise.

System Interface – Home Page:

The above figure shows the main interface of the system where users can perform all features

Fig. 1. Home UserInterface

In this figure, the user interacts with UI

Fig. 2. Analzyed output

Fig. 3. Dashboard

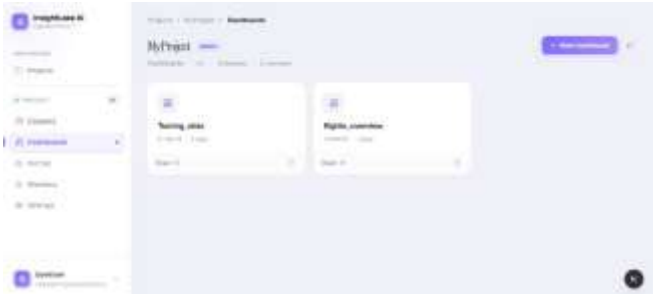

Fig4. Datasets

VI. CONCLUSION

The Insight Lake AI project successfully demonstrates a modern, intelligent, and efficient approach to data analytics by combining the power of Artificial Intelligence with advanced Lakehouse architecture. In today's world, organizations generate huge amounts of data from applications, websites, sensors, business systems, and online platforms. Converting this raw data into useful insights is very important for growth and decision-making. However, traditional analytics systems are often difficult to use because they require technical knowledge, SQL coding skills, and complex tools. This project solves those challenges by providing a simple and user-friendly platform where users can interact with data using natural language queries. One of the major achievements of the system is the AI-based query engine, which understands user questions written in plain English and automatically converts them into SQL queries. This removes the need for manual coding and makes data analysis easier for non-technical users such as managers, students, business teams, and decision-makers. The system also includes query validation and optimization features that check for syntax errors, invalid table names, or wrong column references before execution. This improves accuracy, reduces failures, and ensures better performance. Smart query planning further enhances efficiency by selecting the best execution engine based on data size and query complexity. The integration of multiple high-performance engines such as DuckDB, Polars, and PySpark is another important strength of the project. DuckDB provides fast analytical processing for small and medium datasets, Polars enables high-speed data transformation and DataFrame operations, and PySpark supports distributed processing for very large datasets. Because of this multi-engine design, the platform is flexible, scalable, and suitable for different business environments. Another valuable feature of the project is its visualization capability. Instead of showing only raw tables or text results, the system presents outputs through interactive dashboards, charts, KPI cards, and dynamic filters. Users can easily identify trends, compare values, monitor performance, and explore insights in a visual way. This improves understanding and helps organizations make quick and informed decisions based on real-time data. The simple frontend interface also improves user experience and encourages continuous interaction with the system.

VII. FUTURE SCOPE

The proposed AI-Powered Collaborative Lakehouse Analytics Platform can be further enhanced by integrating real-time data streaming capabilities to support continuous data ingestion and instant analytics. Future improvements may include the incorporation of advanced machine learning and deep learning models for predictive and prescriptive

analytics, enabling the system to provide not only insights but also intelligent recommendations. The platform can also be extended with voice-based query interaction and multilingual support to improve accessibility for a wider range of users. Additionally, deploying the system on cloud infrastructure with containerization and microservices architecture can enhance scalability, reliability, and performance. Integration with advanced data governance, security mechanisms, and role-based access control will further strengthen the system for enterprise-level applications. Collaborative features such as shared dashboards, user roles, and real-time team analytics can also be developed to support organizational decision-making. Overall, these enhancements will transform the system into a more robust, scalable, and intelligent analytics platform suitable for real-world industrial and research applications.

VIII. ACKNOWLEDGMENT

We would like to express our sincere gratitude to our project guide, **Mrs. M. Prasanna Kumari**, Associate Professor, Department of Computer Science and Engineering (Data Science), Vidya Jyothi Institute of Technology, Hyderabad, for his valuable guidance, continuous support, and encouragement throughout the development of this project. His insightful suggestions and motivation greatly contributed to the successful completion of this work.

We would also like to thank the **Head of the Department and faculty members of the CSE (Data Science) department** for providing the necessary support and resources required for carrying out this project. We extend our sincere thanks to the **Principal and management of Vidya Jyothi Institute of Technology** for providing the infrastructure and academic environment that helped us complete this project successfully.

Finally, we express our heartfelt gratitude to our **parents, friends, and well-wishers** for their constant encouragement and support during the course of this work.

IX. REFERENCES

[1] M. Armbrust et al., *"Lakehouse: A New Generation of Open Platforms,"* CIDR, 2021.https://people.eecs.berkeley.edu/~matei/paper s/2021/cidr_lakehouse.pdf

[2] *"Data Lakehouse: A Survey and Experimental Study,"* ScienceDirect, 2024.https://www.sciencedirect.com/science/articl e/pii/S0306437924001182

[3] *"Building an AI-Ready Data Strategy Using Lakehouse Technology,"* 2024. https://alkindipublishers.org/index.php/jcsts/rticle/view/9422

[4] *"AI/ML Optimized Lakehouse Architecture,"* 2025. https://wjaets.com/sites/default/files/fulltext_pdf/WJAETS-2025-0754.pdf

[5] *"The Data Lakehouse: An Evolving Paradigm in Data Architecture,"* 2024. *https://www.researchgate.net/publication/393 759319*

[6] *"Modern Data Lakehouse Architectures,"* 2024. *https://aimjournals.com/index.php/ijaair/artic le/view/466*

[7] *"AI Data Lakehouse Guide,"* Lifebit, 2025. https://lifebit.ai/blog/ai-data-lakehouse-ultimate-guide/

[8] *"Data Lakehouse Architecture Overview,"* 2024. *https://www.scribd.com/document/851955989*

[9] *"Lakehouse Concept Data Architecture,"* 2024. *https://www.lakehousepartners.ai/post/lakeho use-concept-data-architecture*

[10] *"Enhancing Clinical Data Infrastructure for AI Research,"* 2024.https://pmc.ncbi.nlm.nih.gov/articles/P MC12357119/