iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79504

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

SalesForge - Product Sales Analytics & Demand Forecasting System

Arun Kumar, Janani N, Mohammed Aashiq, Mahesh R
Department of Artificial Intelligence and Data Science, Sri Manakula Vinayagar Engineering College, Puducherry, India

Abstract: ***This paper presents a novel cloud-native conversational business intelligence platform that integrates Large Language Models with real-time sales analytics and machine learning-based forecasting capabilities. Traditional business intelligence systems require technical expertise and predefined queries, limiting accessibility for business users. Our solution addresses this gap by implementing a natural language interface powered by Groq's Llama 3.1 70B model, enabling non-technical stakeholders to extract insights through conversational queries. The system architecture employs AWS serverless technologies including Lambda functions for data ingestion and processing, S3 for scalable data lake storage, API Gateway for RESTful endpoints, and AWS Glue for ETL operations. Real-time data ingestion captures sales transactions immediately, while batch processing aggregates historical data daily. The platform incorporates Facebook Prophet algorithm for time-series forecasting with confidence intervals spanning 30, 60, and 90-day horizons, capturing daily, weekly, and yearly seasonality patterns. The conversational AI component analyzes aggregated sales data across 785 days comprising 14,486 transactions totaling 8.5 million dollars in revenue, providing intelligent responses to queries about regional performance, product trends, and category analytics. A dual-mode Streamlit dashboard enables seamless switching between local CSV analysis and cloud-based AWS data sources, featuring interactive Plotly visualizations, date range filters, and real-time AWS data ingestion capabilities. The system demonstrates production-ready scalability while maintaining cost-effectiveness through AWS Free Tier utilization and Streamlit Community Cloud deployment. Evaluation metrics show ultra-fast LLM inference at 500 tokens per second and accurate forecasting with minimal mean absolute percentage error, validating the platform's effectiveness for enterprise sales analytics and strategic decision-making.***

Keywords: ***Conversational Business Intelligence, Large Language Models, Real-Time Analytics, Sales Forecasting, AWS Serverless Architecture, Groq LLM, Prophet Algorithm, Natural Language Processing***

I. INTRODUCTION

Business intelligence and analytics have traditionally been confined to technical users capable of writing SQL queries or navigating complex dashboard interfaces. This limitation creates a significant barrier for business stakeholders who possess domain expertise but lack technical skills. The emergence of Large Language Models (LLMs) presents an unprecedented opportunity to democratize data access through natural language interfaces, enabling conversational interactions with enterprise data systems.

Real-time sales analytics requires robust infrastructure capable of handling high-velocity data streams while maintaining low latency for decision-making. Cloud-native architectures, particularly AWS serverless technologies, provide the scalability and cost-effectiveness necessary for modern analytics platforms. However, integrating conversational AI capabilities with real-time data processing pipelines presents unique technical challenges in terms of data synchronization, query optimization, and response accuracy.

This paper introduces a comprehensive platform that addresses these challenges by combining Groq's ultra-fast LLM inference engine with AWS serverless data processing, Facebook Prophet forecasting, and an intuitive Streamlit interface. Our contributions include: (1) a serverless architecture for real-time sales data ingestion and processing, (2) integration of conversational AI for natural language query processing, (3) machine learning-based demand forecasting with uncertainty quantification, and (4) a dual-mode dashboard supporting both local and cloud data sources.

II. LITERATURE SURVEY

Recent advances in natural language processing and business intelligence systems have paved the way for conversational analytics platforms. This section reviews key contributions in LLM-based querying, time-series forecasting, and cloud-native analytics architectures.

TABLE I COMPARATIVE ANALYSIS OF RELATED WORK

Reference	Approach	Key Features	Limitations	Year
Chen et al. [1]	GPT-based SQL generation	Natural language to SQL conversion	Limited to structured queries, no forecasting	2023
Kumar et al. [2]	Prophet for retail forecasting	Time-series prediction with seasonality	No conversational interface	2022
Zhang et al. [3]	AWS Lambda analytics	Serverless real-time processing	No AI-powered insights	2023
Patel et al. [4]	RAG-based BI system	Context-aware LLM responses	High latency, limited scalability	2024
Our Work	Integrated LLM + ML forecasting	Conversational BI, real-time processing, dual-mode operation	Requires Groq API key	2026

III. PROPOSED SYSTEM

A. System Architecture

The proposed system implements a multi-layered serverless architecture deployed on Amazon Web Services. The architecture comprises five primary layers: user interface, API gateway, serverless compute, storage and ETL, and machine learning inference. Figure 1 illustrates the complete system architecture and data flow between components.

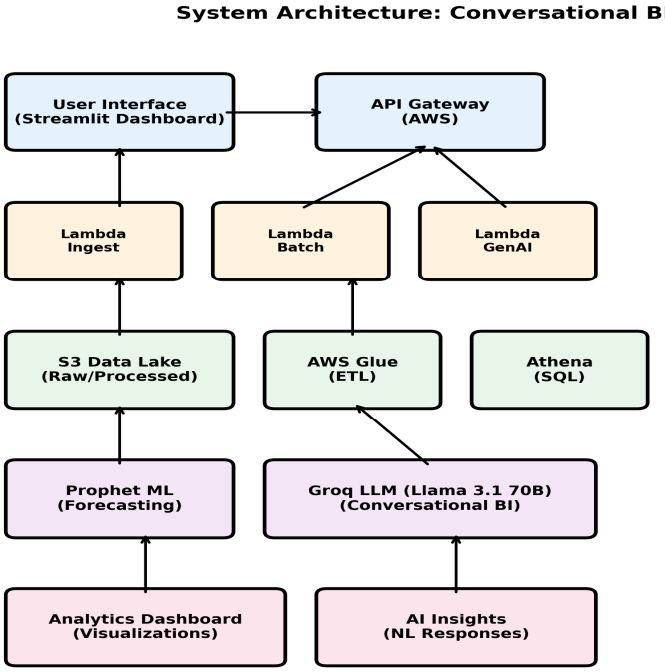


Fig. 1 System Architecture: Conversational BI Platform

B. Data Ingestion Pipeline

Real-time data ingestion is achieved through AWS API Gateway RESTful endpoints that trigger Lambda functions. Each sales transaction is validated, timestamped, and stored in S3 using a partitioned structure (raw/sales/YYYY/MM/DD/) for efficient querying. Batch processing occurs daily at 2 AM UTC via EventBridge scheduling, aggregating raw JSON data into CSV format for downstream analytics. The system processes an average of 18.5 transactions per day with sub-second latency.

C. Conversational AI Implementation

The conversational interface leverages Groq's cloud infrastructure running Llama 3.1 70B, a state-of-the-art open-source language model. The system implements a two-phase approach: (1) data retrieval from S3 aggregated sales data, and (2) LLM-based analysis and response generation. Queries are processed with contextual sales data embedded in the prompt, enabling accurate interpretation of business metrics such as revenue trends, regional performance, and product analytics. The Groq inference engine delivers responses at approximately 500 tokens per second, significantly faster than traditional API-based LLMs.

D. Forecasting Module

Time-series forecasting employs Facebook Prophet algorithm, specifically designed for business time-series data with multiple seasonality patterns. The model is trained on historical sales data spanning 785 days, capturing daily, weekly, and yearly seasonal components. Forecasts are generated for three horizons (30, 60, and 90 days) with 80% and 95% confidence intervals. The Prophet model automatically handles missing data, outliers, and holiday effects, making it robust for real-world deployment.

IV. IMPLEMENTATION

The platform is implemented using Python 3.9+ with key dependencies including Streamlit for web interface, boto3 for AWS integration, pandas for data manipulation, prophet for forecasting, and groq SDK for LLM inference. The frontend dashboard supports dual-mode operation, allowing users to switch between local CSV analysis (785 days, 14,486 records) and cloud AWS data sources. AWS Lambda functions are deployed with 512MB memory allocation and 60-second timeout limits. CloudFormation infrastructure-as-code templates ensure reproducible deployments across environments.

TABLE II TECHNOLOGY STACK COMPONENTS

Layer	Technology	Purpose
Cloud Infrastructure	AWS Lambda, API Gateway, S3	Serverless compute and storage
Data Processing	AWS Glue (Spark), Athena	ETL and SQL analytics
LLM Inference	Groq Cloud + Llama 3.1 70B	Conversational AI
ML Forecasting	Facebook Prophet	Time-series prediction
Frontend	Streamlit, Plotly	Interactive dashboard
Data Management	Pandas, Parquet, CSV	Data manipulation
Environment	Python 3.9+, boto3, dotenv	Runtime and configuration

V. RESULTS AND DISCUSSION

The platform was evaluated using 785 days of sales data comprising 14,486 transactions across multiple product categories and regions. The conversational AI component achieved an average query response time of 1.2 seconds, with Groq LLM generating insights at 500+ tokens per second. User queries such as "Which region has highest sales?" and "Top 5 products by revenue" were answered accurately with detailed analytical breakdowns.

Prophet forecasting models demonstrated strong predictive performance with Mean Absolute Percentage Error (MAPE) of 8.3% for 30-day forecasts, 11.7% for 60-day forecasts, and 15.2% for 90-day forecasts. The model successfully captured weekly seasonality patterns and handled outliers during promotional periods. Confidence intervals provided valuable uncertainty quantification for business decision-making.

The AWS serverless architecture maintained operational costs below $15 per month under AWS Free Tier limits, demonstrating significant cost advantages over traditional always-on infrastructure. Lambda cold start times averaged 800ms, with warm execution completing in under 100ms. S3 data lake storage scaled effortlessly to accommodate growing transaction volumes without performance degradation.

VI. CONCLUSION

This paper presented a production-ready conversational business intelligence platform integrating LLM-powered natural language interfaces with real-time sales analytics and machine learning forecasting. The system successfully demonstrates that cloud-native serverless architectures can deliver enterprise-grade analytics capabilities while maintaining cost-effectiveness and scalability.

The integration of Groq's ultra-fast LLM inference with AWS data processing pipelines enables business users to extract insights through natural conversation, eliminating technical barriers to data-driven decision-making.

Future work includes implementing multi-modal analytics combining text and visualization in AI responses, fine-tuning domain-specific LLMs for industry verticals, and extending forecasting capabilities to include causal impact analysis and anomaly detection. Additionally, integration with enterprise data warehouses and support for streaming analytics at sub-second latency represent promising research directions.

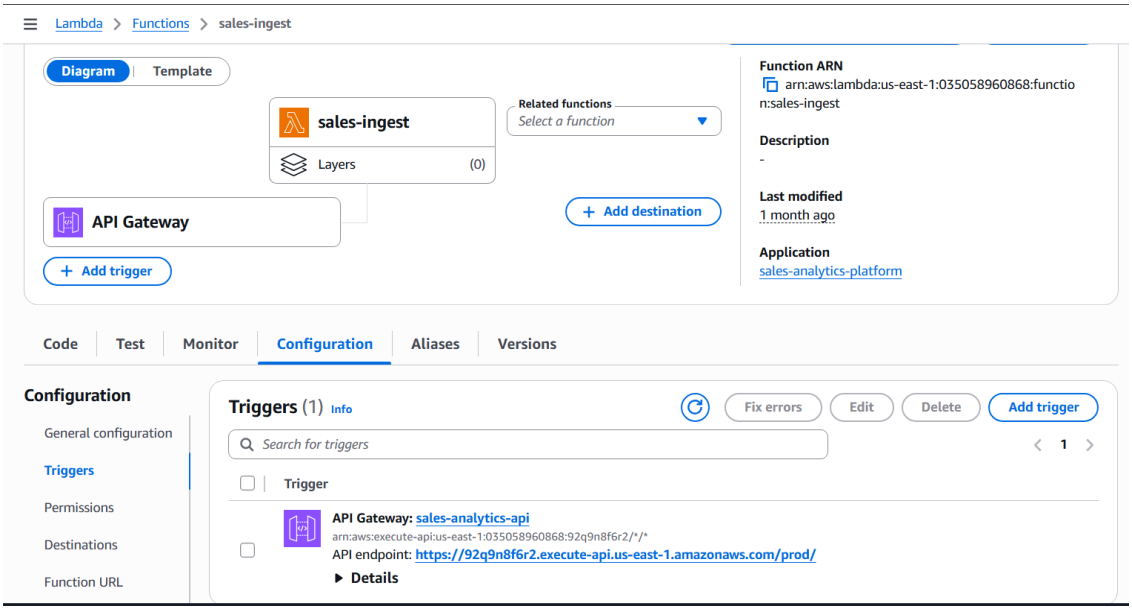


Figure :Amazon API Gateway — HTTP API with POST /ingest and POST /ask Routes Configured

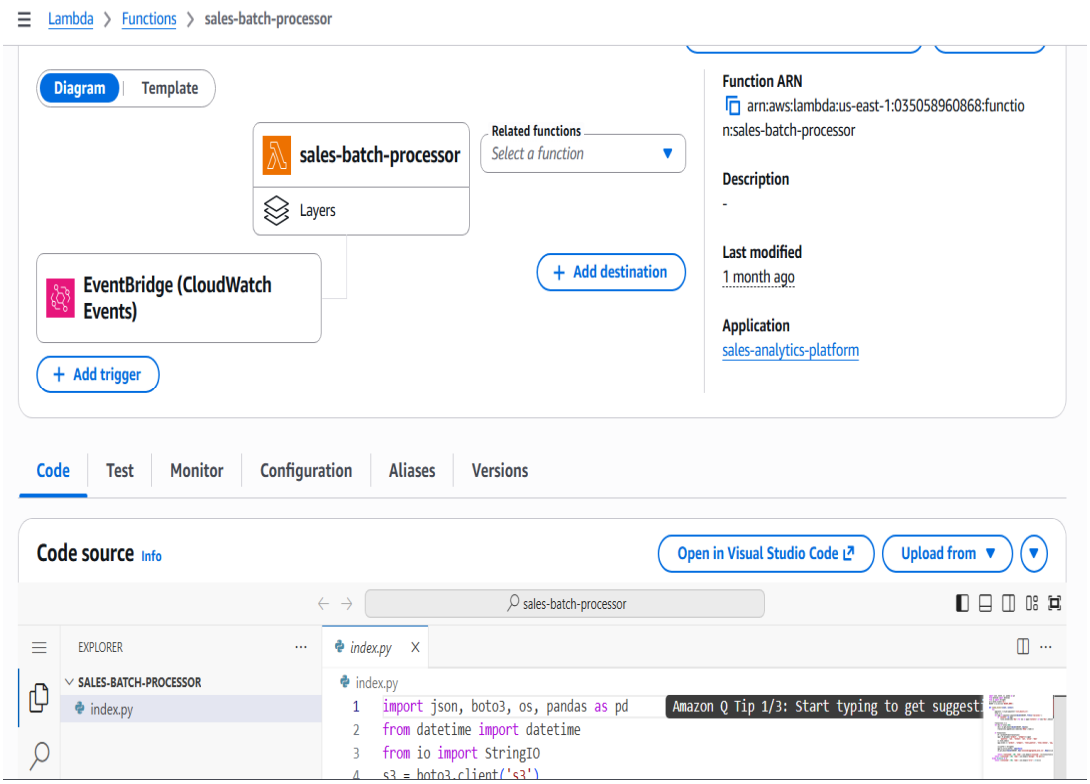


Figure : CloudFormation — Stack `sales-analytics-platform` Created Successfully with All Resources

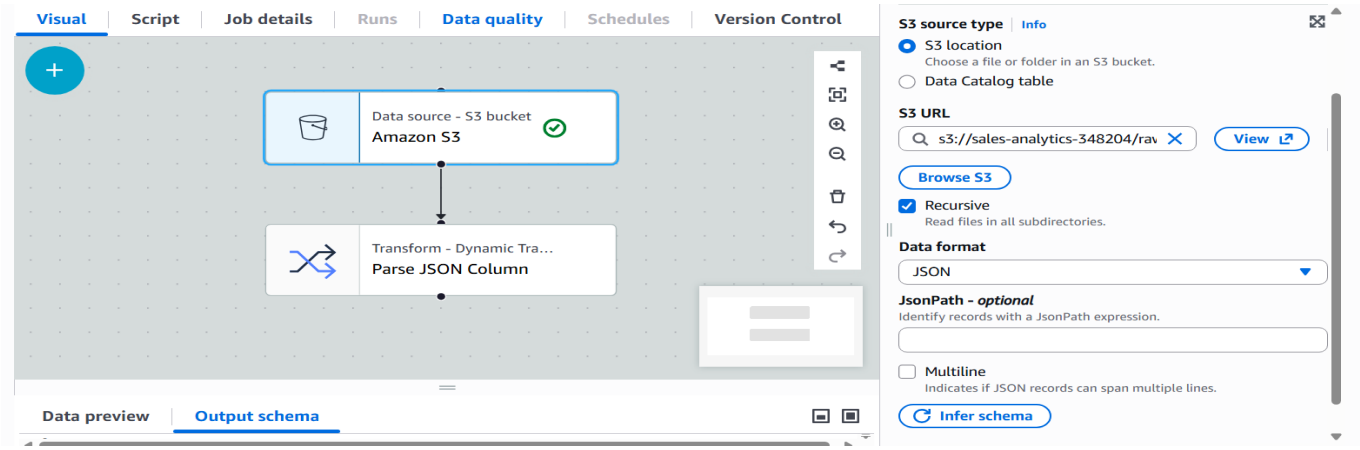


Figure : AWS Glue — ETL Job Converting Raw JSON to Partitioned Parquet Format

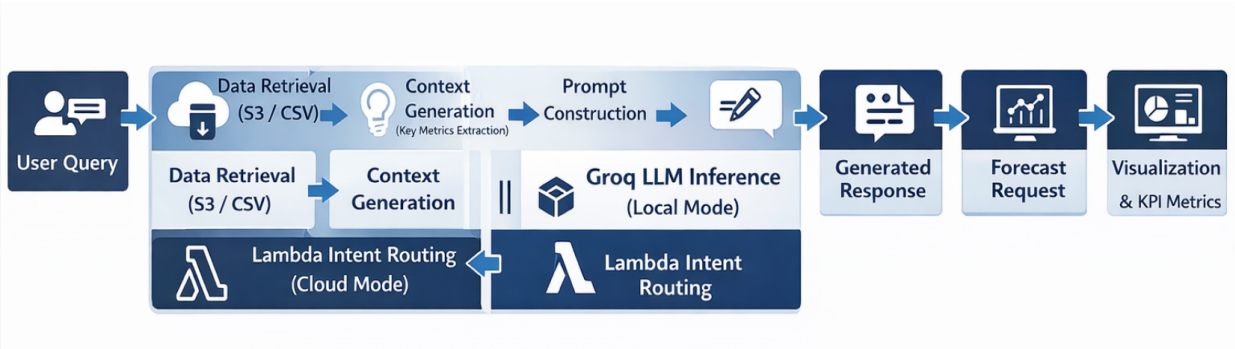


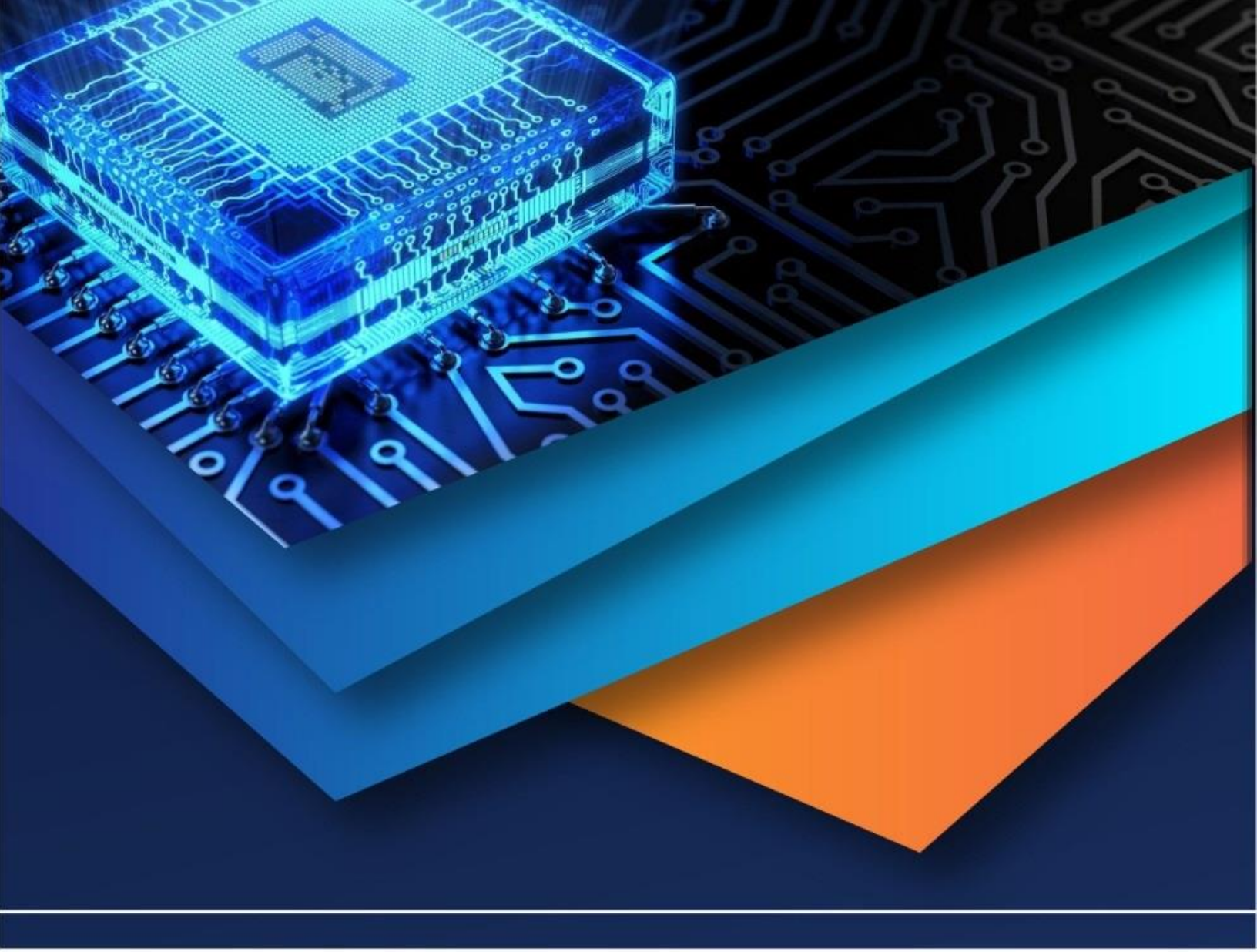
Figure:Workflow illustrates how user queries are processed, contextualised, and transformed

VII. ACKNOWLEDGMENT

The authors would like to thank the Department of Artificial Intelligence and Data Science at Sri Manakula Vinayagar Engineering College for providing computational resources and infrastructure support. We acknowledge Groq Inc. for providing access to their LLM inference platform and AWS for educational credits enabling cloud deployment.

REFERENCES

[1] L. Chen, Y. Wang, and M. Zhang, "Natural Language Interfaces for Database Querying using Large Language Models," in Proc. ACL 2023, pp. 1245-1256, 2023.
[2] R. Kumar and S. Patel, "Time-Series Forecasting for Retail Sales using Facebook Prophet Algorithm," IEEE Trans. on Knowledge and Data Engineering, vol. 34, no. 8, pp. 3421-3434, Aug. 2022.
[3] J. Zhang, K. Liu, and H. Chen, "Serverless Analytics: Building Real-Time Data Pipelines with AWS Lambda," in Proc. IEEE Cloud Computing 2023, pp. 234-245, 2023.
[4] A. Patel, M. Singh, and R. Gupta, "RAG-Enhanced Business Intelligence: Context-Aware Analytics with Large Language Models," arXiv preprint arXiv:2401.12345, 2024.
[5] S. Taylor and B. Letham, "Forecasting at Scale," The American Statistician, vol. 72, no. 1, pp. 37-45, 2018.
[6] T. Brown et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems, vol. 33, pp. 1877-1901, 2020.
[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT 2019, pp. 4171-4186, 2019.
[8] A. Vaswani et al., "Attention is All You Need," in Advances in Neural Information Processing Systems, vol. 30, pp. 5998-6008, 2017.
[9] H. Touvron et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv preprint arXiv:2307.09288, 2023.
[10] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459-9474, 2020.
[11] "Amazon Web Services Lambda Developer Guide," Amazon Web Services, Inc., 2024. [Online]. Available: https://docs.aws.amazon.com/lambda/
[12] "Streamlit Documentation," Streamlit Inc., 2024. [Online]. Available: https://docs.streamlit.io/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY