Leveraging AI/LLM Models for Real-Time Analytics in Online Applications

Jawaharbabu Jeyaraman*, Kumarasenthil Muthuvel. (* *indicate presenting author*)

Amtech Analytics

Abstract

Application logs, typically semi-structured, present a significant challenge when real-time insights are required. Extracting meaningful data from logs often involves reliance on developers to manually mine and analyze the logs, or requires centralizing, formatting, and structuring the logs before insights can be generated. This manual process can lead to delays and inefficiencies in diagnosing application issues.

Our solution leverages Large Language Models (LLMs) to automatically format, structure, and analyze log data, enabling real-time insights into the application's performance and issues. By integrating logs generated from Node.js (PM2 servers) into a real-time pipeline using Amazon CloudWatch, ElasticSearch, and Pinecone (Vector Database), our system transforms semi-structured log data into actionable knowledge. The LLM models, deployed via Amazon Bedrock, enable users to interact with logs through a Slack-based chatbot, providing immediate insights without the need for manual intervention from developers.

This AI-driven approach not only streamlines log management but also democratizes access to application performance data, empowering non-technical users to troubleshoot issues effectively and efficiently. Our solution sets the stage for future innovations, including predictive analytics and automated anomaly detection.

Problem Statement

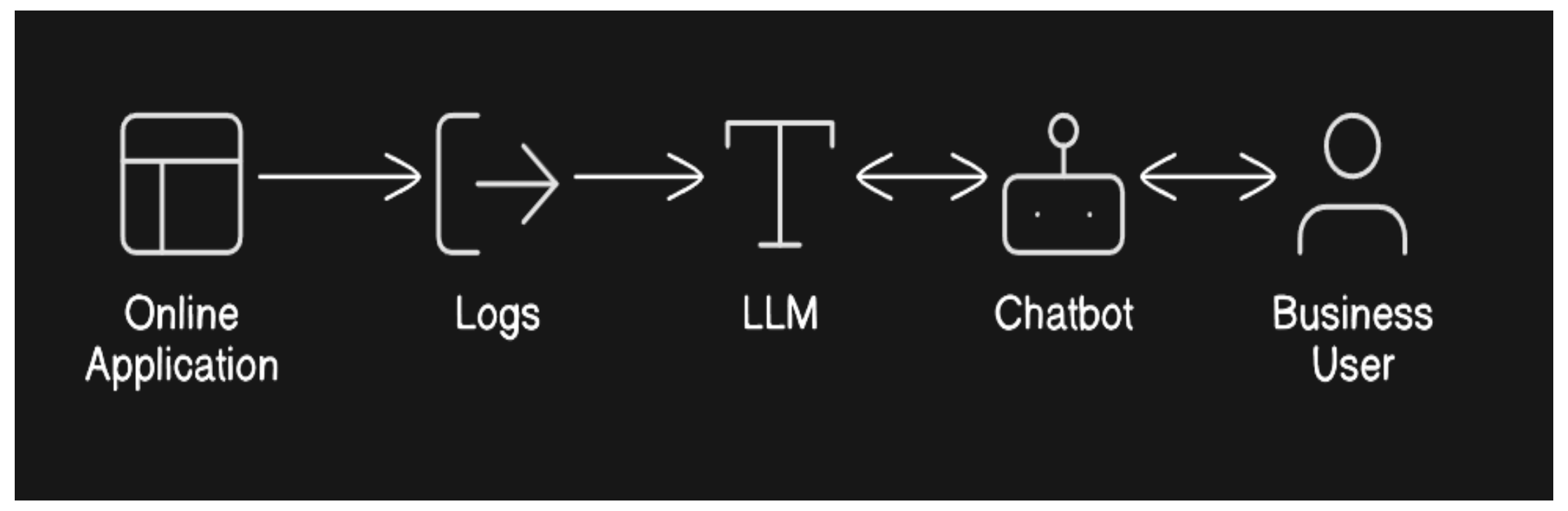
Problem Statement: Application logs are often semi-structured, making it difficult to gain real-time insights into the performance and issues of an application. Traditionally, developers are needed to manually mine the logs, format, and analyze them to provide relevant insights. This process is slow and inefficient, requiring significant manual intervention or complex centralized systems for log management.

Solution: Large Language Models (LLMs) can automatically format, structure, and interpret log data. By using LLMs, we can convert semi-structured logs into organized, structured data, enabling real-time insights into application performance without relying on manual processing.

Background and Method of Approach

Solution Overview

In our proposed solution, we address the challenge of extracting real-time insights from semi-structured application logs by leveraging the power of Large Language Models (LLMs). The system takes application logs generated in real time, formats and structures them using LLMs, and stores the resulting structured data in a dynamic knowledge base. This enables non-technical users to query logs through natural language and receive immediate insights on various application stages and errors.



High-Level Solution Flow

The diagram illustrates the core components of the proposed solution to address the challenge of gaining real-time insights from semi-structured application logs. The flow of the solution is as follows:

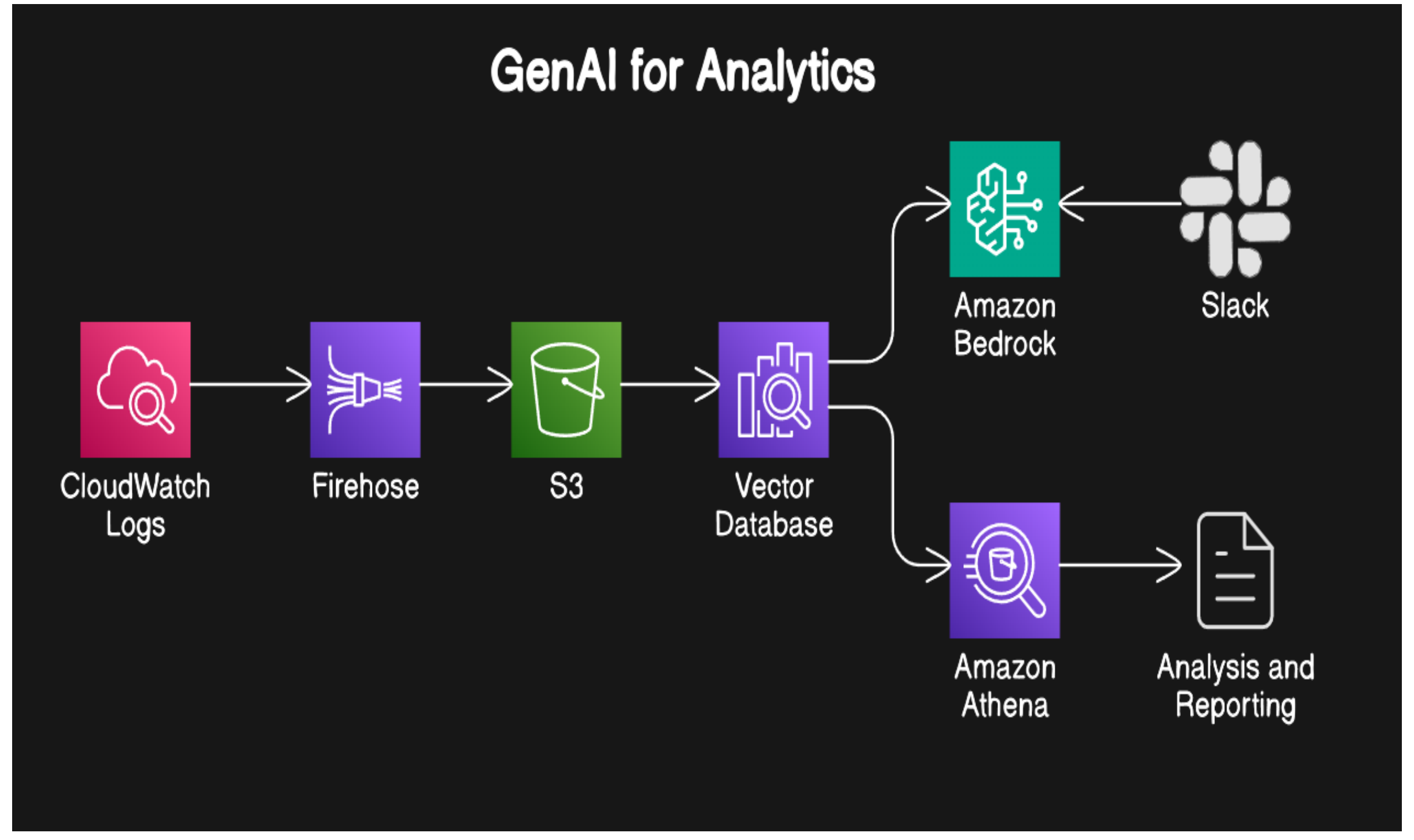
1. **Online Application**: Logs are generated in real time from the online loan vending application. These logs contain critical data regarding application performance, stages, and any encountered errors.
2. **Logs**: The logs are transmitted to a central logging system (such as Amazon CloudWatch). These logs are semi-structured, requiring further processing to be transformed into actionable insights.
3. **LLM (Large Language Model)**: The logs are processed by an LLM, which automatically formats and structures the semi-structured log data. The LLM helps convert raw log entries into structured information that can provide insights into application stages, errors, and performance metrics.
4. **Chatbot Interface**: The LLM is integrated with a chatbot (such as a Slack-based chatbot) that enables users to interact with the processed log data in real time. Business users can query the chatbot using natural language, receiving structured insights from the logs.
5. **Business User**: The final output is delivered to business users, allowing them to query and gain real-time insights into the application logs without relying on developers or technical expertise. They can identify application issues, track stages, and understand overall performance directly through the chatbot interface.

This solution empowers users by streamlining the process of log analysis, transforming raw logs into useful insights through automation with LLMs and delivering those insights via a user-friendly chatbot.

System Architecture

Log Generation & Storage:
- **Node.js** (PM2 server) generates logs for each stage of the loan vending application.
- Logs are forwarded to **Amazon CloudWatch Logs** for centralized monitoring.

Real-Time Data Pipeline:
- Logs are retrieved in real-time from CloudWatch and processed.
- Data is indexed and stored in either **ElasticSearch** (for structured querying) or **Pinecone Vector Database** (for unstructured, semantic queries), with **Amazon S3** as the primary storage for raw log files.
- The logs act as a **knowledge base**, continuously updating with application activity.


GenAI for Analytics

Integration with AI/LLMs:
- **LLM Models** (via **Amazon Bedrock**) are integrated to enhance query capabilities.
- Logs are analyzed using standard LLM models that are fine-tuned to understand specific application terms, logs, and error codes.
- AI models provide intelligent responses based on log data, enabling users to query specific stages of the application or identify key errors.

User Interface & Chatbot:
- **Slack Integration**: Users interact with the system via a Slack chatbot, which provides real-time responses to log queries.
- The chatbot is built using **Amazon Bedrock**, allowing users to query the knowledge base using natural language, making it easy for non-technical users to retrieve valuable insights.
- The chatbot can return specific application states, stages, and errors upon request, offering actionable insights into performance and issues.

Results

Key Benefits

- **Real-Time Error Tracking**: Immediate identification of application issues at any stage, reducing downtime and improving the customer experience.
- **Enhanced Decision-Making**: The integration of LLMs allows for deeper insights into logs and error patterns, making it easier for stakeholders to make informed decisions.
- **Scalable & Flexible**: The use of **ElasticSearch** and **Pinecone** ensures the solution can handle large-scale data and complex queries with ease, scaling as the application grows.
- Analysis & Reporting: Use of vector Database also serves the purpose of real time analysis and report generation and real time querying.
- **Seamless Integration**: By utilizing **Slack** and **Amazon Bedrock**, users can easily access log information without needing technical expertise.

Innovation & Future Work

- **LLM Model Fine-Tuning**: Future iterations will focus on fine-tuning models to automatically detect and predict issues before they escalate, using historical data from the knowledge base.
- **Broader Application Scope**: This system can be extended to other domains where log analysis and real-time issue resolution are critical, such as healthcare, e-commerce, and IoT systems.
- **Enhanced Analytics**: Introduce predictive analytics and anomaly detection to provide proactive recommendations to users.

Conclusion

Incorporating AI and LLM models into the loan vending application analytics pipeline significantly enhances error tracking, operational efficiency, and user experience. By creating a robust, AI-powered log management system, we enable users to resolve issues in real-time, improve application performance, and ensure smoother loan processing.

Real Time Communications Conference and Expo at Illinois Tech
IEEE International Conference
IEEE RTC 2024 - October 8-10, 2024 - IIT, Chicago, IL, USA