

# Gen AI IT Log Analyzer | Enable IT Teams to Chat with Infrastructure on Dell PowerEdge™ R760xa Server with Nvidia H100 Data Center Tensor Core GPUs.

| July 2024

Retrieval Augmented Generation IT Log Analysis on Dell PowerEdge™ R760xa Server with Nvidia H100 Data Center Tensor Core GPUs.



The exponential growth of machine data from modern IT infrastructure, including logs from applications, servers, and network devices, presents a significant challenge for IT teams. While the large volume and complexity of the data makes it difficult for IT teams to manually analyze the logs to identify critical errors and anomalies in a timely manner, traditional log analysis methods are also labor-intensive and often fall short in providing timely, accurate, and correlated insights from different log streams, leading to potential system downtime, vulnerabilities, and inefficiencies.

Imagine an enterprise workflow in which the infrastructure communicates problems to IT Teams instead of the opposite, eliminating the need for extensive manual RAG log analysis while streamlining the error identification process.



Using retrieval-augmented generation (RAG), infrastructures can be enabled to provide detailed insights into system and network issues, allowing IT Teams to diagnose and resolve them efficiently. Dell, in partnership with Nvidia and Metrum AI, is thrilled to unveil a cutting-edge GenAI-Based IT Log Analyzer Solution that leverages Generative AI and RAG to revolutionize the log analysis and incident management experience and process.

### **This solution enables the following key advancements in IT Log Analysis:**

- **Efficiency and Accuracy:** By enabling AI-assisted or fully automated log analysis, the solution significantly reduces the time and effort required to identify and diagnose issues, ensuring more accurate and timely insights.
- **Scalability:** The solution can handle large volumes of log data from multiple sources, scaling seamlessly with the growing needs of IT infrastructures.
- **AI-based Correlation:** By correlating errors and identifying patterns, the solution can help in early diagnosis thereby avoiding potential system downtime, security threats and vulnerabilities more effectively.
- **Enhanced User Experience:** A user-friendly interface enables AI-assisted diagnosis of error and source log reference through chat conversation. The added capability to generate detailed RCA reports and incident tickets further enhances the user experience.

In this blog, Metrum AI walks through an enterprise-ready solution architecture and provides insights into the user interface along with example user queries.

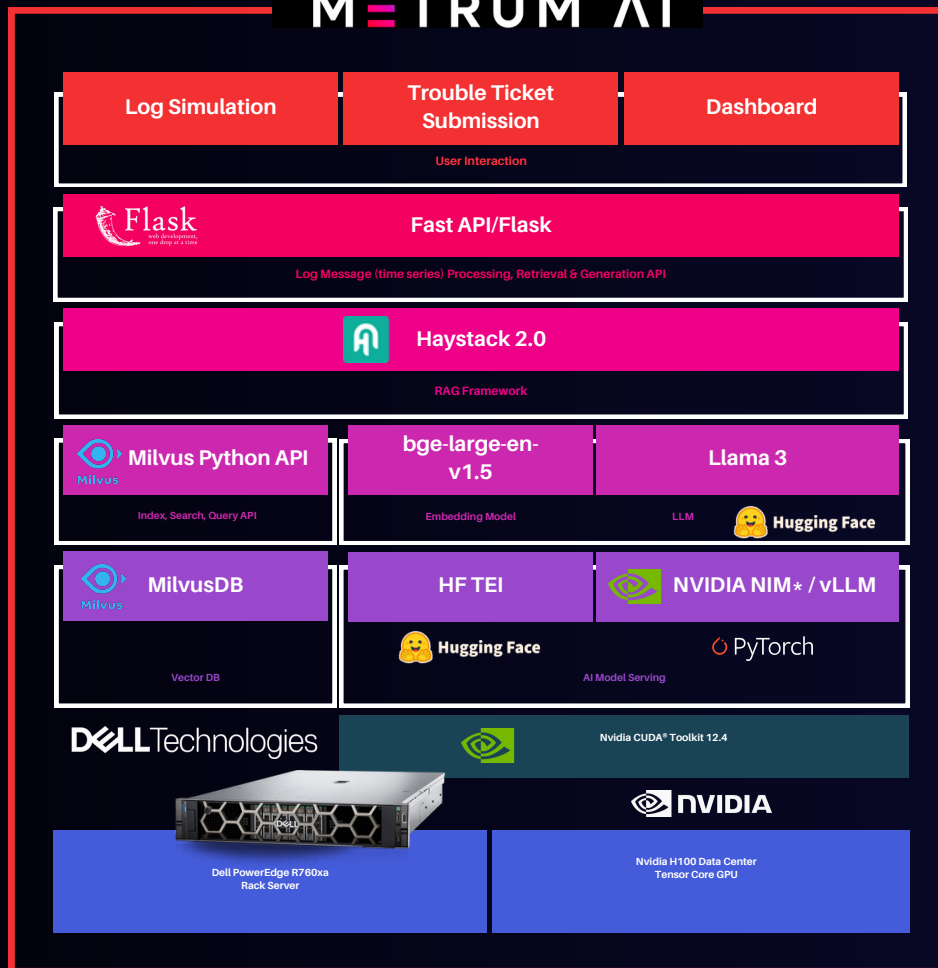
### **Solution Architecture**

This solution leverages Dell PowerEdge R760xa server equipped with Nvidia H100 Data Center Tensor Core GPUs, and utilizes a suite of models, including bge-large-en-v1.5, a text embedding model, as well as the Llama 3 large language model. This solution is built on Haystack 2.0, an industry leading RAG framework, along with Nvidia NIM, offering optimized microservices at scale.

**The image below illustrates the solution architecture in detail.**

See diagram on next page.

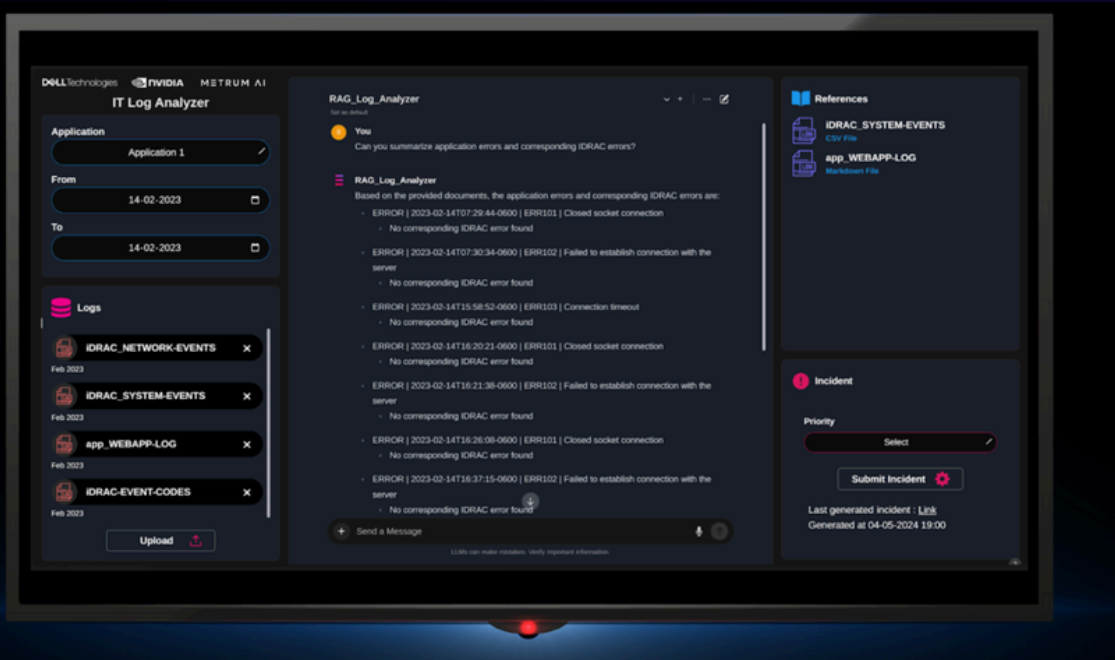
# METRUM AI



## Key Features

- **Log Ingestion Interface:** This solution enables easy ingestion from multiple log sources including applications, servers, and networks.
- **Customizable Log Analysis:** Users can seamlessly select target applications and specific time frames for analysis.
- **Chat Console:** The solution enables users to converse with the application and query information relevant to uploaded files.
- **Error Identification and Retrieval:** This solution automatically detects errors, provides detailed insights and references to the source logs, and enhances troubleshooting without manual intervention.
- **Intelligent Error Correlation:** The application correlates errors across multiple sources, paving the way for early diagnosis and preventing potential downtime.
- **Automated Incident Ticket Generation:** Following the conversation and analysis, the system auto-generates an incident ticket in a PDF format, capturing all relevant details for record-keeping and further action.

## Step-by-Step Demo Walkthrough



1. The user ingests logs from different devices and types such as application, server and network through the user interface.
2. The user selects the application and time frame for log analysis.
3. The user converses with the application through the conversation console.
4. The application retrieves and automatically provides identified errors, along with their details and source log references.
5. The application automatically correlates errors across multiple sources and helps in the early diagnosis of errors.
6. The application automatically allows users to generate an incident ticket (PDF) based on the analysis results.

## Example Scenarios

The following example scenarios simulate a user query relevant to the uploaded IT log files, along with the corresponding text response, reference list, and incident ticket generated by the application.

Can you summarize application errors from 7:00 to 17:00 as a markdown table?

The screenshot shows the IT Log Analyzer interface. The central chat area displays the following content:

**RAG\_Log\_Analyzer**

You: Can you summarize application errors from 7:00 to 17:00 as a markdown table?

RAG\_Log\_Analyzer: Here is the summary of application errors from 7:00 to 17:00 as a markdown table:

Timestamp	Error Code	Error Message
2023-02-14T07:29:44-0600	ERR101	Closed socket connection
2023-02-14T07:30:34-0600	ERR102	Failed to establish connection with the server
2023-02-14T16:20:21-0600	ERR101	Closed socket connection
2023-02-14T16:21:38-0600	ERR102	Failed to establish connection with the server
2023-02-14T16:26:08-0600	ERR101	Closed socket connection
2023-02-14T16:36:05-0600	ERR101	Closed socket connection
2023-02-14T16:37:15-0600	ERR102	Failed to establish connection with the server
2023-02-14T16:46:58-0600	ERR101	Closed socket connection
2023-02-14T16:55:55-0600	ERR104	API request failed

Note: There are 9 application errors between 7:00 and 17:00.

Can you summarize application errors and corresponding IDRAC errors from the same time period?

The screenshot shows the IT Log Analyzer interface. The central chat area displays the following content:

**RAG\_Log\_Analyzer**

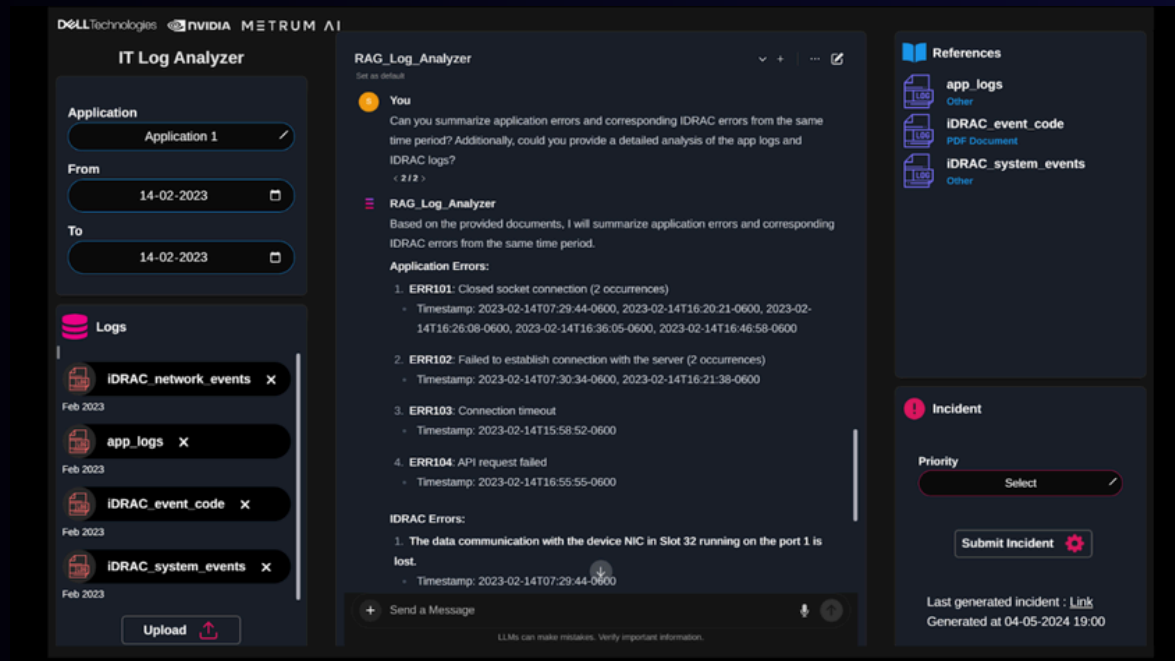
You: Can you summarize application errors and corresponding IDRAC errors from the same time period?

RAG\_Log\_Analyzer: Based on the provided documents, here is a summary of application errors and corresponding IDRAC errors from the same time period:

- ERR101: Closed socket connection** at 2023-02-14T07:29:44-0600
  - Corresponding IDRAC error: **The data communication with the device NIC in Slot 32 running on the port 1 is lost.** at 2023-02-14T07:29:44-0600
- ERR102: Failed to establish connection with the server** at 2023-02-14T07:30:34-0600
  - No corresponding IDRAC error found.
- ERR103: Connection timeout** at 2023-02-14T15:58:52-0600
  - No corresponding IDRAC error found.
- ERR104: API request failed** at 2023-02-14T16:46:58-0600
  - No corresponding IDRAC error found.

Note: There are no other application errors or IDRAC errors that match the same timestamp.

Can you summarize errors and corresponding IDRAC errors from the same time period? Additionally, could you provide a detailed analysis of the app logs and IDRAC logs?



The image below details an example GenAI incident report.

**Date and Time of Incident:** 2023-02-14

**Incident Summary**

The application encountered errors related to socket connections, failed API requests, and memory allocation. The IDRAC logs indicate warning messages related to data communication loss with the device NIC and unable to log in for root from 10.134.139.221 using GUI.

**Incident Details**

- Application Name:** Web server
- Error Message(s):**
  - ERR101: Closed socket connection
  - ERR102: Failed to establish connection with the server
  - ERR103: Connection timeout
  - ERR104: API request failed
  - ERR107: Memory allocation failed
- Affected Functionality:** API requests, socket connections, and memory allocation
- Number of Users Affected:** N/A (not specified)

**Investigation Details**

- Reviewed server logs: The server logs show a series of errors related to socket connections, failed API requests, and memory allocation.
- Checked recent code changes: No recent code changes were found that could be related to the errors.
- Examined user reports: No user reports were found that could be related to the errors.

**Correlated Issues:**

- Error 101: Closed socket connection:** This error occurred multiple times, indicating issues with socket connections. The IDRAC logs show warning messages related to data communication loss with the device NIC, which may be related to this error.
- Error 102: Failed to establish connection with the server:** This error occurred when the application tried to establish a connection with the server. The IDRAC logs show warning messages related to unable to log in for root from 10.134.139.221 using GUI, which may be related to this error.
- Error 103: Connection timeout:** This error occurred when the application timed out while waiting for a response from the server. The IDRAC logs show warning messages related to data communication loss with the device NIC, which may be related to this error.
- Error 104: API request failed:** This error occurred when the application failed to make an API request. The IDRAC logs show warning messages related to unable to log in for root from 10.134.139.221 using GUI, which may be related to this error.
- Error 107: Memory allocation failed:** This error occurred when the application failed to allocate memory. The IDRAC logs show a critical event related to a memory device at A7, which may be related to this error.

It appears that the application errors are related to issues with socket connections, failed API requests, and memory allocation, which may be caused by issues with the IDRAC system or the device NIC. Further investigation is needed to determine the root cause of the errors and to resolve the issues.

**Reference Log Files:**

- app\_WEBAPP-LOG.md
- IDRAC\_NETWORK-EVENTS.csv
- IDRACEventcode.pdf

## Summary

Dell PowerEdge R760xa server equipped with Nvidia H100 Data Center Tensor Core GPUs, offers enterprises industry-leading infrastructure to create custom RAG solutions using their proprietary data. In this blog, we showcased how enterprises deploying applied AI can take advantage of RAG capabilities in the context of an IT Log Analyzer, uncovering the following milestones:

- Built end-to-end RAG on Dell PowerEdge R760xa server with Nvidia H100 Data Center Tensor Core GPUs, validated on Nvidia NIMs, and deployed with Haystack RAG Framework.
- Enabled IT professionals to chat with network, application, and infrastructure logs leading to faster root cause analysis and auto generated reports.
- Showcased live at Dell Technologies World '24.

## Additional Criteria for IT Decision Makers:

### What is RAG, and why is it critical for enterprises?

RAG, which stands for Retrieval-Augmented Generation, is a method in natural language processing (NLP) that enhances the generation of responses or information by incorporating external knowledge retrieved from a large corpus or database. This approach combines the strengths of retrieval-based models and generative models to provide more accurate, informative, and contextually relevant outputs.

The key advantage of RAG is that it leverages a large amount of external knowledge dynamically, enabling the model to generate responses that are not just based on its training data but also on up-to-date and detailed information from the retrieval phase. This makes RAG particularly useful in applications where factual accuracy and detail are crucial, such as in customer support, academic research, and other domains requiring precise information. Ultimately, RAG provides enterprises with a robust tool for improving the accuracy, relevance, and efficiency of their information systems, leading to better customer service, cost savings, and competitive advantages.

## Resources

Dell product images: [Dell.com](https://www.dell.com)

Copyright © 2024 Metrum AI, Inc. All Rights Reserved. This project was commissioned by Dell Technologies. Dell and other trademarks are trademarks of Dell Inc. or its subsidiaries. Nvidia and combinations thereof are trademarks of Nvidia. All other product names are the trademarks of their respective owners.

\*\*\*DISCLAIMER - Performance varies by hardware and software configurations, including testing conditions, system settings, application complexity, the quantity of data, batch sizes, software versions, libraries used, and other factors. The results of performance testing provided are intended for informational purposes only and should not be considered as a guarantee of actual performance.