CASE STUDY

How Pluton Biosciences Leveraged Retrieval-Augmented Generation (RAG) to Revolutionize Microbial Research with AI

Executive Summary

Pluton Biosciences, a biotech firm in St. Louis, Missouri, specializes in computational biology to uncover microbes for agriculture and climate solutions. They partnered with us to develop a Retrieval-Augmented Generation (RAG) model that bridges the gap between large language models (LLMs) and domain-specific expertise. Leveraging AWS architectures, Bedrock, Terraform, and advanced AI, we created an automated data pipeline integrating lab-generated and scholarly data. This scalable, cost-efficient solution uses Amazon Titan and Anthropic's Claude models to minimize hallucinations and deliver highly accurate results. The project significantly enhanced Pluton's research efficiency, marking a successful collaboration in the pursuit of groundbreaking scientific advancements.

Customer Description

The Pluton Biosciences team brings together scientists, entrepreneurs, technologists, and an advisory board of accomplished scientists with a shared passion for protecting our climate, our planet and the people who call it home.

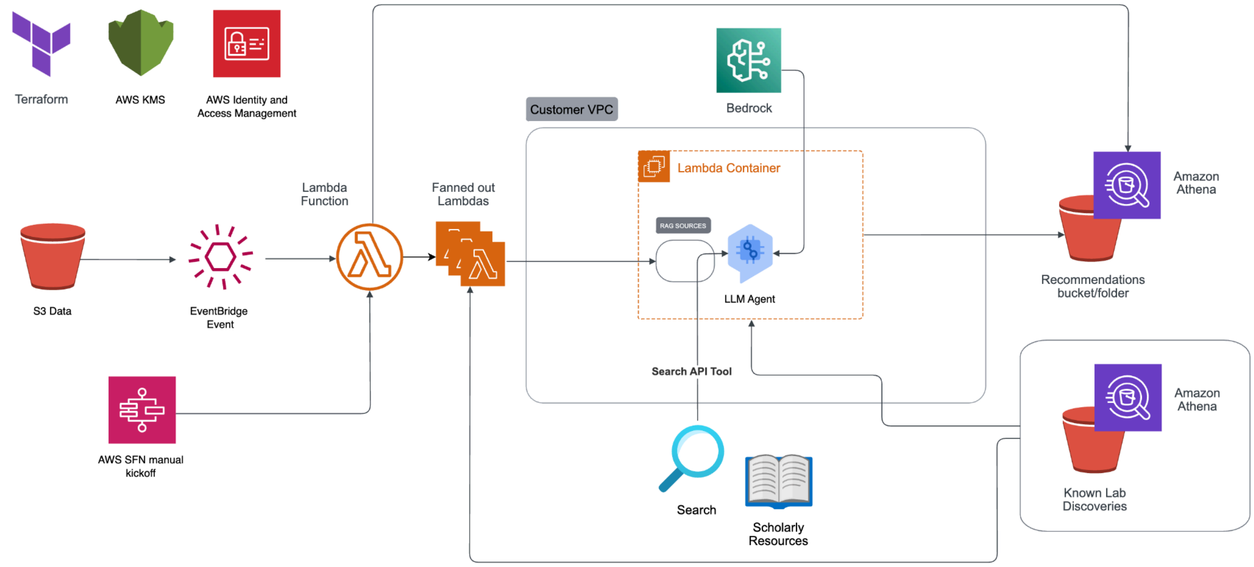
Description of Service

Recently, we partnered with our friends at Pluton Biosciences to develop a Retrieval-Augmented Generation (RAG) model that enhances their research capabilities. Pluton, a biotech firm located in St. Louis, Missouri, extensively utilizes computational biology to explore and analyze new microbes that can be applied in agriculture and climate solutions.

Large language models (LLMs) alone frequently produce inaccuracies and lack sufficient domain expertise, making them unreliable for accelerating research at Pluton. RAG addresses the disconnect between the linguistic capabilities of LLMs and the essential domain knowledge that a well-designed retrieval system can provide. By integrating this model with streamlined retrieval and indexing of pertinent data, we empowered Pluton to enhance its research efficiency at a minimal cost per run.

The project was built by combining Pluton's deep knowledge of microbiology and computational biology, our deep understanding of AWS architectures, Bedrock, Terraform, and AI, and some shared love of biological sciences. We built this as an automated data pipeline that integrates with their work and calls an LLM agent, augmented with appropriate data. This process receives data generated from the lab and vectorized relevant scholarly data. It uses both Amazon's Titan and Anthropic's Claude models to produce a highly accurate RAG model for identifying key features in the organisms relevant to their groundbreaking work.

The architecture for the project looks as follows:



The architecture can handle increased demand as Pluton grows because the process scales efficiently using fanned-out lambdas based on the number of samples received. No architectural changes are needed to support this scalability.

It is important to note that we reduced hallucinations - an issue many GenAI models encounter - by ensuring that the model returns the cited output in a consistent format.

Description of Outcome

Overall, this project was a great success. We loved working with Pluton Biosciences and encourage you to check out some of their amazing work! If you are interested in us helping your company build an AI-driven solution, please reach out! At New Math Data we are an AWS Generative AI Competency Partner and are both skilled in and passionate about our work.