rapyder
Solutions, Simplified.

Rapyder empowers Stellaps with a cutting-edge Gen-AI powered multilingual chatbot!

stellapps
Smart Systems, Stellar Applications.

Introduction

Stellapps stands out as the first-of-its-kind startup in India dedicated to digitizing the dairy supply chain. Founded in 2011, it is an I IT Madras-incubated, Bangalore-based Internet of Things {IoT) startup with a primary focus on data acquisition and machine learning. Milk, the largest crop on this planet, highlights a strong demand for technology interventions, particularly in emerging markets where the yield per animal is low, traceability is inadequate, and quality is not up to the mark.

Business Need

StellApps has an application the caters to farmers from rural India. The application helps farmers to get the detail for the farming related queries. Currently the existing application is not able to handle the native language features because of that application is not helping farmers who are the end-user of the application. The StellApps team wants to enhance the application by introducing a feature that allows users to record audio questions in their native language. These questions relate to topics such as taking care of cattle, managing/handling the crops etc. In response, the application should provide response in

AWS Services

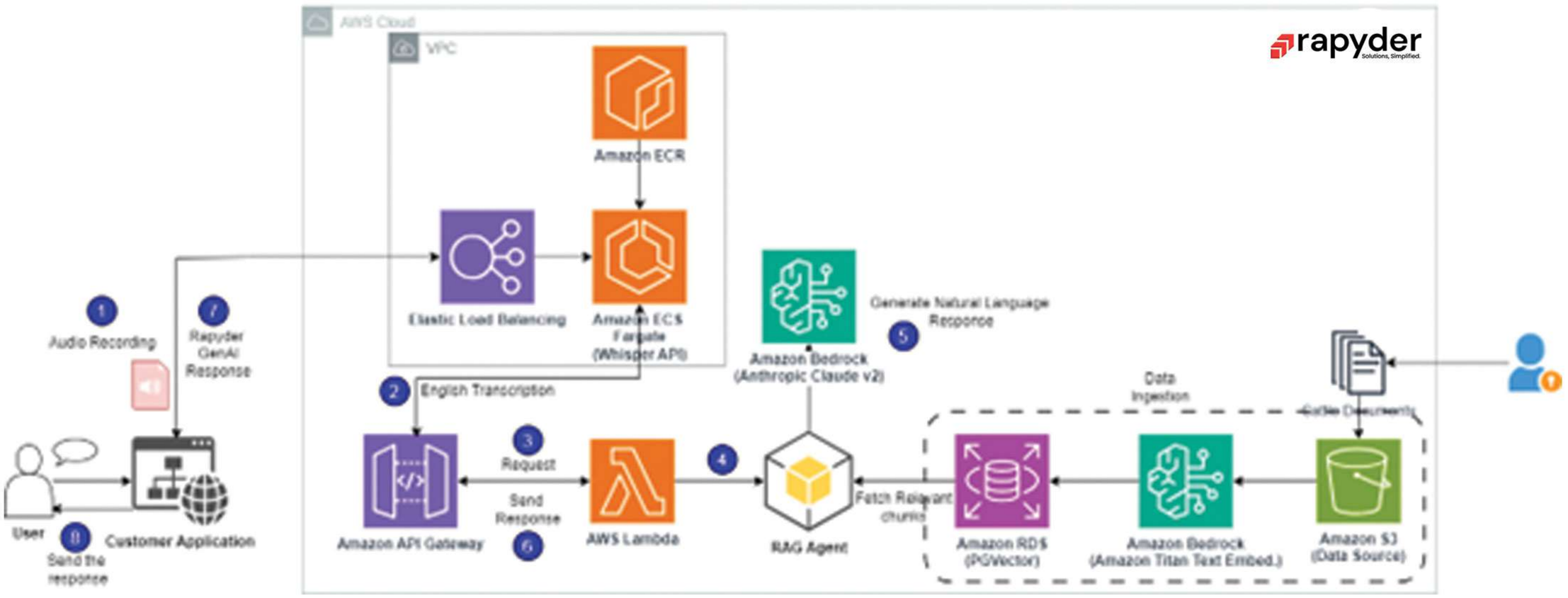
AWS VPC, NAT Gateway, Lambda, ECS with Fargate, ECR, AWS Bedrock, AWS RDS, API Gateway, S3

Industry

IoT

Solution Architecture

The core of our GenAI solution is a microservices-based architecture deployed on the AWS cloud. This architecture provides scalability, flexibility, and resilience, ensuring that the chatbot can meet the demands of Stellapps.



Proposed Solution:

- As per AWS best practices, we have kept raw data in S3 bucket. An AWS Lambda event will be triggered to pre-process the uploaded data.
- We are performing the embedding using Amazon Bedrock Titan model on the raw data, and ingesting the embeddings into Amazon RDS PGVector DB.
- We have created Whisper ECS API that accepts audio, and performs the transcription, and parse transcription to LLM API.
- We are converting the English input text into embeddings using Amazon Bedrock Titan model.
- We have implemented the data retrieval on VectorDB using similarity search algorithm using Langchain based approach.
- Response is getting created by using Claude2 model with help of Langchain based RAG pipeline using refined prompts after getting input from the StellApps team.
- As part of best practices, if the context or answer is not available for the given query chatbot is politely asking the end-user to connect with customer care and relevant field expert doctors.
- Used chunking for the knowledge-based data while embedding data into PGVector.
- Experimented different “K” values while setting up RAG pipeline to get the better results while performing similarity for the query.
- Performed parameter tuning for TopP, token length and temperature control for getting the good accuracy.
- Added prompt engineering best practices and ensured that no false information is being provided as response and input is added to get the information vetted from the doctor or field

Reaping Rewards

- Added feature to handle Indian languages with higher accuracy, this will help to get a larger Indian user-base.
- Reduced effort and maintenance cost for Knowledge base pipeline.
- Better response time resulted in satisfied customer and smooth user experience.

rapyder
Solutions, Simplified.
+91 733 868 6644
info@rapyder.com
www.rapyder.com