Kalpa Optimise Proposal

Description

- Kalpa Optimise is an AI-powered cloud optimisation platform designed to reduce AWS cloud expenditure. It provides recommendations to efficiently manage cloud expenditures and enable better scaling decisions.
- The platform leverages AI to analyse past and real-time cloud usage and expenditure data from CUR and CloudWatch, fine-tuned using LLMs, to provide actionable insights for cost reduction and dynamic scaling.

Problem Statement

Organizations often struggle with:

- Managing Cloud Costs: Due to the dynamic nature of usage and the wide range of pricing models available.
- Resource Inefficiency: Leading to wasted expenditure.
- Operational Disruptions: Caused by unexpected usage patterns.
- Scaling Challenges: Difficulty in scaling resources effectively to meet demand.

This results in:

- Unpredictable costs
- Inability to optimize across cloud Architecture
- A lot of time is spent on manual oversight
- Increased risk of resource wastage, a poorly performing AWS cloud model with high cost.

Solution Overview

Kalpa Optimise addresses these issues by providing an AI-powered platform that automates:

- AWS cloud architecture optimization
- Dynamic scaling recommendations.
- The platform analyzes past and real-time cloud usage and expenditure data from CUR and CloudWatch.
- It leverages fine-tuned LLMs to provide recommendations for:

This results in:

- Efficient resource recommendation.
- Cost-effective cloud architecture decisions.
- Scaling actions to match resource allocation to demand.
- Alerts for unusual usage patterns that may indicate overutilization & wastage.

AI Necessity

AI is crucial for Kalpa Optimise because it enables:

- Automated analysis of large volumes of cost and usage data from CUR and CloudWatch.

- Dynamic recommendations for resource optimization and scaling actions using fine-tuned LLMs.
- AI provides unique value by automating complex decision-making processes and delivering actionable insights for cost savings, proactive issue identification, and efficient resource management.

AWS Architecture:
https://drive.google.com/drive/folders/1nq2N-Qu3kJFIJ25N7Wc8dOh9zkW94yNm?usp=sharing

AI Methodology

- Utilizing data from CUR and CloudWatch, tagging recommendations and anomalies to the available data.
- Fine-tuning LLMs using the SageMaker / Bedrock platform to generate cost optimization, scaling, and anomaly detection insights.

AI Implementation Rationale

The choice of AI is justified by the need to:

- Process and analyze complex cloud cost and usage data efficiently from CUR and CloudWatch.
- Automate cost optimization and scaling recommendations and reduce the need for manual intervention, using fine-tuned LLMs.
- Deliver precise and actionable recommendations and alerts, maximizing cost savings and minimizing operational risks.

Key AWS Services

1. **AWS SageMaker / AWS Bedrock:** To train and deploy the fine-tuned LLM & To register and manage the fine-tuned LLM for generating recommendations and alerts.
2. **AWS CloudWatch:** To fetch operational metrics, logs, and insights from the client into AWS S3.
3. **AWS S3:** Secure storage for uploading AWS Cost and Usage Reports.
4. **AWS RDS for Postgres:** Database for storing transformed data and AI model outputs.
5. **AWS Glue:** For running the ETL pipeline.

Kalpa Optimise has the potential to significantly impact organizations by:

- Reducing AWS cloud expenditure.
- Improving cloud cost management efficiency.
- Enabling better decision-making regarding cloud resource allocation.
- Optimizing resource utilization and performance through intelligent scaling recommendations.

The platform is designed to be scalable by leveraging AWS services like S3, RDS, Glue, API Gateway, Amazon Bedrock & AWS SageMaker, which can handle increasing data volumes, user demand, and real-time data processing needs.