OpsVoice Assistant: Executive Summary

Submission for Google Cloud x Datadog AI Partner Catalyst

The Problem: The "Black Box" of AI Latency

As enterprises race to adopt Generative AI, SREs (Site Reliability Engineers) face a new observability gap. When an AI-powered application is slow, traditional monitoring tools cannot distinguish between network lag, database locks, or LLM inference delays. This "blind spot" leads to wasted cloud spend and poor user retention.

The Solution: OpsVoice Assistant

OpsVoice is a voice-activated observability tool that integrates **Google Vertex AI (Gemini Pro)** with **Datadog APM** to provide end-to-end visibility into AI pipelines.

- **Voice-to-Insight:** Engineers can ask questions naturally via voice commands.
- **Granular Tracing:** We implemented custom Datadog Spans to separate Speech-to-Text latency (Application Layer) from Vertex AI inference latency (Model Layer).
- **Real-Time Dashboards:** Visualizes "Tokens per Second" and "Model Latency" alongside standard infrastructure metrics.

Key Technical Achievements

1. **Distributed Tracing:** Successfully instrumented a Python/Flask backend to trace requests across Google Cloud services.
2. **Root Cause Identification:** In live demonstrations, OpsVoice detected a **33-second latency spike** and correctly identified the root cause as the Gemini model inference (vertex-ai-gemini) rather than the transcription service (google-stt), which took only 102ms.
3. **Serverless Architecture:** Fully containerized using **Docker** and deployed on **Google Cloud Run** for auto-scaling and zero-maintenance overhead.

Tech Stack

- **Compute:** Google Cloud Run (Serverless)
- **AI Models:** Google Vertex AI (Gemini Pro), Google Speech-to-Text
- **Observability:** Datadog APM (ddtrace), DogStatsD
- **Language:** Python 3.11 (Flask, Gunicorn)

Market Potential

With the AI Observability market projected to reach **$20B+ by 2028**, OpsVoice addresses a critical immediate need for the 40% of enterprises currently deploying LLMs in production.