

TIER 1 & TIER 2 · AI MIGRATION & USE CASE DISCOVERY

AI Use Case Discovery Assessment

A technical and financial roadmap for migrating a production AI workspace platform to AWS — evaluating cost optimization, model economics, and a phased path to a lower-cost, higher-margin architecture.

8/10

TECHNICAL READINESS

3

PRIORITY USE CASES

85%

PROJECTED COST
REDUCTION

212%

PROJECTED 12-MONTH
ROI

PREPARED FOR

**AI-Powered Technical Workspace
Platform — Document Generation,
Diagramming & Research Tools**

ASSESSMENT PERIOD

June 23 – June 25, 2024

ASSESSMENT TYPE

**Tier 1 — AI Production Workload
Migration Assessment, with Tier 2 — Use
Case Discovery**

PREPARED BY

Flentas Technologies Pvt. Ltd.

REPORT DATE

June 26, 2024

CLOUD PLATFORM

Amazon Web Services (AWS)

— EXECUTIVE SUMMARY

1

The Case for Migrating to AWS

IN SHORT

The platform's core costs are driven by third-party LLM providers running on Google Cloud, at roughly \$7,000 a month. Migrating the workload to an AWS-native stack built around Amazon Nova is projected to cut that cost by approximately 85% while establishing a standardized foundation for the company's other AI products in development.

1.1 Assessment Objectives

This assessment was conducted to perform a comprehensive evaluation of the Client's flagship AI workspace platform, an existing AI application currently in production on Google Cloud Platform (GCP). The objectives were to analyze the current architecture, performance, and operational costs associated with third-party LLM providers; evaluate the feasibility and benefits of migrating the entire workload to Amazon Web Services (AWS); and develop a detailed implementation roadmap and ROI analysis for the migration.

1.2 Key Findings

Strategic opportunities identified

PRIORITY USE CASE 1

Platform Migration & Cost Optimization**Highest Impact**

The platform's core features are powered by expensive third-party LLMs, primarily Anthropic models, resulting in high operational costs of roughly \$7,000 per month. Migrating the entire application from GCP to an integrated AWS stack, centered on cost-effective models like Amazon Nova, presents the single largest opportunity for reaching profitable scale.

PRIORITY USE CASE 2

High Value

Standardize Future Development on AWS

The Client is actively developing other AI-powered tools, including an SEO platform and a talent-matching engine. Migrating the flagship platform would establish a scalable, secure, and cost-optimized AWS architecture that can serve as a standardized foundation for these future products, accelerating their time to market.

PRIORITY USE CASE 3

Forward-Looking

Enhance Platform Capabilities

An integrated AWS stack opens the door to enhancing the platform with additional capabilities, such as advanced observability with Amazon CloudWatch, intelligent search with Amazon Kendra for the retrieval pipeline, and more robust CI/CD and MLOps practices.

Technical Readiness

Overall Readiness Score

8 / 10

KEY STRENGTHS

The platform is 100% cloud-native, built on a sophisticated RAG-based, multi-LLM architecture. The in-house technical team is highly capable, having built and launched a successful beta to more than 500 users.

CRITICAL GAPS

The primary constraint on growth is the high and unpredictable monthly cost of the current LLM providers. The platform has no direct integration with AWS Bedrock, and the team lacks a model-routing layer, which adds friction to switching between providers.

Recommended AWS AI Solution

Primary model recommendation: a comprehensive evaluation and subsequent adoption of Amazon Nova models to replace the current Anthropic and OpenAI models.

Estimated implementation timeline: six months for a phased migration.

Expected ROI: approximately 212% within twelve months.

1.3 Business Impact Summary

METRIC	CURRENT STATE (MONTHLY)	TARGET STATE (MONTHLY)	IMPROVEMENT
LLM operational cost	~\$7,000+	~\$1,050 (projected with Nova)	~85% reduction
Platform scalability	Limited by cost	Unlocked for 5,000–10,000 users	Significant growth
Cross-platform development time	Not applicable	Reduced with a standardized stack	Efficiency gain

1.4 Immediate Next Steps

- ✓ **Develop an AWS Bedrock integration:** build the necessary integration to allow for model evaluation. Target completion: July 19, 2024.
- ✓ **Conduct a hands-on model evaluation workshop:** run a direct comparison of Amazon Nova models against the current Anthropic model for the core document-generation use case. Target completion: July 26, 2024.
- ✓ **Finalize the migration plan and SOW:** based on the workshop results, create a detailed project plan and Statement of Work for the phased migration to AWS. Target completion: August 9, 2024.

METHODOLOGY

2

Assessment Overview

2.1 Assessment Methodology

This assessment was conducted following AWS Well-Architected AI Lens principles and the AWS RAPID AI Pilot program requirements. The methodology included:

- Use case discovery:** a detailed product demonstration and stakeholder interviews to understand the platform, its business objectives, and its customer value proposition.
- Technical assessment:** a thorough evaluation of the existing GCP architecture, data flows, security posture, and third-party integrations to identify migration pathways and potential challenges.
- Model evaluation:** a discussion-based analysis of the current LLM providers, Anthropic and OpenAI, and a proposed framework for benchmarking them against Amazon Nova models on performance, cost, and feature set.

- Implementation planning:** the design of a high-level technical architecture on AWS and development of a phased implementation roadmap.
- Business case development:** a detailed ROI analysis focused on the financial benefits of migrating to a more cost-effective, AWS-native AI stack.

2.2 Stakeholders Engaged



Ashish Malav

Co-Founder, Client organization



Manav

Technical Lead, Client organization



Kapil

Client organization



Ganga

Flentas Technologies



Rushabh

Flentas Technologies

— OPPORTUNITY MAPPING

3

Use Case Discovery & Prioritization

3.1 Business Context Analysis



Strategic Objectives

Provide a comprehensive, AI-powered workspace that consolidates the entire workflow for technical teams — from research and brainstorming to generating complex technical documents such as BRDs, PRDs, and architecture diagrams — replacing the inefficiency of multiple disparate tools.



Current Challenges

The platform's primary challenge is the high and unpredictable monthly cost of its primary LLM provider, which limits profitable scaling. The company is also seeking a long-term, scalable cloud partner for this and future AI products.



Success Criteria

Maintaining or improving the quality and consistency of AI-generated content while significantly reducing operational costs. Scaling the user base to 5,000–10,000 users profitably is a key business goal.

3.2 Identified Use Cases

PRIORITY USE CASE 1

AI-Powered Technical Document Generation

Medium Complexity

DESCRIPTION

The platform's core feature. Users input a brief concept, and the platform generates a complete, well-structured technical document, such as a BRD or SOW, using a series of clarifying questions and a sophisticated prompt-engineering framework.

AI APPLICATION

Currently uses Anthropic's Claude models via Google Vertex AI. The system is tuned using a RAG-based approach to produce domain-specific, high-quality output.

PRIORITY USE CASE 2

Automated Diagram Generation

Medium Complexity

DESCRIPTION

An extension of the document studio in which the platform reads a generated BRD and automatically creates a corresponding user workflow or architecture diagram.

AI APPLICATION

This feature uses an LLM to interpret the semantic structure of the document and translate it into a visual flow, saving significant manual effort.

PRIORITY USE CASE 3

Integrated Research Chatbot

Medium Complexity

DESCRIPTION

A conversational agent within the workspace that performs in-depth research on behalf of the user to support the early phases of product development.

AI APPLICATION

The chatbot uses a deep-reasoning model to run multiple, iterative web searches and compile comprehensive research on a given topic, tailored to the user's professional profile.

3.3 Use Case Prioritization Matrix

USE CASE	BUSINESS IMPACT	IMPLEMENTATION EFFORT	TECHNICAL FEASIBILITY	PRIORITY
Document Generation	High	Medium	High	1
Research Chatbot	High	Medium	High	2
Diagram Generation	Medium	Medium	High	3

4

— INFRASTRUCTURE & SECURITY

Technical Readiness Assessment

4.1 Overall Technical Readiness: 8 / 10

KEY STRENGTHS

- Cloud-native foundation:** the entire application is built on modern, cloud-native principles, currently running on GCP Cloud Run.
- Advanced AI architecture:** a sophisticated multi-LLM, RAG-based system with vector embeddings (PGVector) delivers context-aware, domain-specific results.
- Experienced in-house team:** the team has demonstrated the capability to build, deploy, and manage complex AI applications from the ground up.
- Market traction:** the platform has successfully launched a beta program, onboarding hundreds of users and demonstrating product-market fit.

CRITICAL GAPS

- Cost management:** current LLM spending is high, at roughly \$7,000 per month, and is the single biggest barrier to profitable growth.
- Vendor lock-in risk:** the current architecture is tightly coupled to GCP services and a single primary LLM provider.
- Lack of AWS integration:** the platform has no existing integration with AWS Bedrock, requiring an estimated three-week development effort to enable model testing.

4.2 Compliance and Security Assessment



Regulatory Requirements

The team is actively working toward GDPR compliance. No specific data residency requirements were identified, but this should remain a consideration for a global platform.



Security Posture

Authentication is handled through Google Authentication, and all customer data is encrypted to protect privacy.



Integration Capabilities

The system is designed for integrations, with plans to connect to Jira, Google Workspace, and Zoho.

4.3 Technical Architecture Recommendations

The proposed AWS architecture would involve:

- **Compute:** migrating the front-end and back-end services from GCP Cloud Run to AWS EC2 or AWS Fargate for scalable, containerized hosting.
- **AI/ML:** replacing the Google Vertex AI endpoint with Amazon Bedrock to access foundation models such as Amazon Nova and Titan.
- **Database:** assessing migration of the current Supabase (Postgres) database, including the PGVector embeddings, to Amazon RDS for PostgreSQL or Amazon Aurora to maintain data continuity.
- **Security and networking:** using AWS IAM for access control and Amazon VPC to create a secure, isolated network environment.

— MODEL BENCHMARKING

5

AWS Model Evaluation Results

5.1 Model Evaluation Framework

A hands-on evaluation could not be performed during the assessment period, due to the lack of an existing AWS Bedrock integration. A framework for a future model comparison workshop was established instead, based on the following criteria.



Performance Metrics

Response quality and consistency, the primary factor for the Client, along with accuracy and latency.



Cost Analysis

Cost per one million tokens (input and output), and total monthly operational cost at the projected scale.



Feature Comparison

Maximum token output, a key requirement for generating long documents, along with multilingual support and integration ease.

5.2 Model Comparison

The following comparison is based on the assessment discussion and public benchmarks.

MODEL	PROJECTED COST (VS. CURRENT)	KEY STRENGTHS	RECOMMENDATION
Amazon Nova Pro	~85% cheaper	Performance on par with GPT-4, a large context window, and optimization for complex document analysis.	Primary recommendation for the document and diagram studios.
Anthropic Claude 3 (current)	Baseline (~\$7,000/mo)	Delivers highly consistent, high-quality results with a large token output.	Use as the performance benchmark for evaluation.
OpenAI GPT-4 (current)	High	Used for some platform features today.	Evaluate Nova for specific research or reasoning tasks.

5.3 Model Recommendations

PRIMARY RECOMMENDATION — AMAZON NOVA

Amazon Nova is projected to deliver performance highly comparable to the current models at a significantly lower price point, estimated at seven to ten times cheaper. This directly addresses the primary business challenge of high operational costs and is the most critical step toward reaching profitable scale.

IMPLEMENTATION

Conduct a hands-on workshop to validate these projections using real-world prompts drawn from the platform's actual production traffic.

6

— DELIVERY PLAN

Implementation Roadmap

6.1 Implementation Strategy

Approach: a three-phase migration starting with the highest-impact, lowest-effort component — integrating AWS Bedrock to validate the cost-saving hypothesis — followed by a feature-by-feature migration and finally a full infrastructure lift and shift.

Timeline: six months for a full platform migration.

Resource requirements: a dedicated project team consisting of the Client's developers and an AWS-certified architect from Flentas.

6.2 Phase-by-Phase Implementation Plan

1

Bedrock Integration & Model Validation

Months 1-2

OBJECTIVES

Establish connectivity to AWS Bedrock and conduct a rigorous, hands-on evaluation of Amazon Nova models against the incumbent Anthropic models.

KEY ACTIVITIES

- Develop an integration layer for Bedrock, potentially via a model-routing layer or direct SDK
- Set up core AWS services, including IAM and VPC
- Execute benchmark tests for the core document-generation use case

DELIVERABLES

A functional Bedrock integration, and a comprehensive model evaluation report with data-driven recommendations.

2**Core Product PoC on AWS**

Months 3–4

OBJECTIVES

Migrate the highest-impact feature, the document studio, to run entirely on AWS using the model selected in Phase 1.

KEY ACTIVITIES

- Refactor the backend to replace GCP and Vertex AI API calls with Amazon Bedrock API calls
- Deploy the feature to a containerized environment on AWS, such as App Runner
- Conduct A/B testing with a segment of beta users to confirm quality and performance parity

DELIVERABLES

A production-ready document studio running on AWS, and a performance and cost report comparing the new stack to the old.

3**Full Platform Migration & Optimization**

Months 5–6

OBJECTIVES

Migrate all remaining features, including the diagram studio and chatbot, along with the underlying infrastructure, database, and front-end hosting, to AWS.

KEY ACTIVITIES

- Migrate the front end from GCP Cloud Run to AWS App Runner
- Execute a database migration from Supabase to Amazon RDS or Aurora
- Implement comprehensive logging, monitoring, and cost-optimization strategies in AWS

DELIVERABLES

The full platform operating on AWS, along with a final project close-out and cost analysis report.

7

— PATH FORWARD

Recommendations & Next Steps

7.1 Strategic Recommendations

PRIMARY RECOMMENDATION

Prioritize the AWS Bedrock model evaluation. The most significant business value lies in reducing the platform's monthly LLM expenditure, and a successful evaluation of Amazon Nova will provide the financial justification for the full migration and unlock a path to profitable scaling.

Expected Outcomes

- A clear, data-driven decision on the foundational model for all future development on the platform
- A potential monthly cost reduction of more than \$5,000

7.2 Immediate Next Steps (30 Days)

- Executive Decision and Approval**
 - Formally approve the three-week development effort to build the AWS Bedrock integration
 - Schedule the hands-on model comparison workshop
- Team Formation and Planning**
 - Allocate engineering resources for the integration development
 - Flentas to assign an AWS solutions architect to support the process
- Technical Preparation**
 - Procure AWS account credentials and configure the necessary IAM permissions
 - Compile a representative dataset of prompts and expected outputs from the platform to use as the benchmark standard in the workshop



GET IN TOUCH

Let's discuss your AI roadmap

Have questions about this assessment or ready to scope the next phase? The Flentas team is available to walk through the findings, refine the migration plan, or kick off the Bedrock integration.



EMAIL

sales@flentas.com



LINKEDIN

[Flentas Technologies](#)

8/10

TECHNICAL
READINESS

85%

PROJECTED COST
REDUCTION

212%

PROJECTED 12-
MONTH ROI

6 mo

MIGRATION
TIMELINE