VERTEX AI IMPLEMENTATION PLAN

Certify Intel - Competitive Intelligence Platform

Version	v5.3.0-VERTEX
Status	PROPOSED (Pending Approval)
Date	January 26, 2026
Estimated Effort	6-8 weeks across 5 phases
Total Tasks	30 tasks
New Code	~6,200 lines across 12 files
Estimated Cost	~$78/month

Executive Summary

This plan outlines the integration of Google Cloud Vertex AI into Certify Intel to enhance competitive intelligence capabilities with enterprise-grade AI features. The migration from the consumer Google AI SDK (google-generativeai) to Vertex AI will unlock:

- **Enterprise Security:** VPC Service Controls, CMEK encryption, HIPAA compliance
- **RAG Engine:** Grounded responses from competitor knowledge bases
- **Agent Builder:** Autonomous competitive intelligence agents
- **Vector Search:** Semantic search across competitor data
- **Model Fine-Tuning:** Custom models trained on healthcare competitive intelligence
- **Multi-Agent Systems:** Coordinated agents for research, analysis, and reporting

Current State vs. Proposed State

Feature	Current (Google AI SDK)	Proposed (Vertex AI)
RAG Engine	Manual implementation	Managed per-competitor corpora
Vector Search	Keyword-based only	Semantic search with embeddings
Agent Builder	Manual orchestration	Autonomous agents with MCP
Fine-Tuning	Not available	Custom CI model training
HIPAA Compliance	Not available	Enterprise BAA available
Security	API key authentication	VPC-SC, CMEK, IAM, audit logs
Grounding	Limited	Full Google Search + corpus grounding

Implementation Phases

Phase 1: Core Vertex AI Migration (Week 1-2)

ID	Task	Priority
VERTEX-1.1	Set up GCP project with Vertex AI	HIGH
VERTEX-1.2	Create vertex_ai_provider.py (~800 lines)	HIGH
VERTEX-1.3	Migrate existing AI calls	HIGH
VERTEX-1.4	Add service account authentication	HIGH
VERTEX-1.5	Update .env configuration	HIGH
VERTEX-1.6	Create provider abstraction	MEDIUM

Phase 2: RAG Engine Integration (Week 2-3)

ID	Task	Priority
VERTEX-2.1	Create RAG corpus management	HIGH
VERTEX-2.2	Build document ingestion pipeline	HIGH
VERTEX-2.3	Implement grounded generation	HIGH
VERTEX-2.4	Add RAG API endpoints	HIGH
VERTEX-2.5	Integrate with battlecard generator	MEDIUM
VERTEX-2.6	Add citation extraction	MEDIUM

Phase 3: Vector Search Implementation (Week 3-4)

ID	Task	Priority
VERTEX-3.1	Create Vector Search index	HIGH
VERTEX-3.2	Build embedding pipeline	HIGH
VERTEX-3.3	Implement semantic search API	HIGH
VERTEX-3.4	Add similarity search	MEDIUM
VERTEX-3.5	Create search UI component	MEDIUM
VERTEX-3.6	Index historical data	LOW

Phase 4: Agent Builder Integration (Week 4-6)

ID	Task	Priority
VERTEX-4.1	Create CI Agent definition	HIGH
VERTEX-4.2	Build MCP tool integrations	HIGH
VERTEX-4.3	Implement agent memory	HIGH
VERTEX-4.4	Add scheduled agent tasks	MEDIUM
VERTEX-4.5	Create agent chat UI	MEDIUM
VERTEX-4.6	Build alert system	MEDIUM

Phase 5: Fine-Tuning & Security (Week 6-8)

ID	Task	Priority
VERTEX-5.1	Prepare fine-tuning dataset	MEDIUM
VERTEX-5.2	Train custom CI model	MEDIUM
VERTEX-5.3	Configure VPC-SC	HIGH
VERTEX-5.4	Set up CMEK	MEDIUM
VERTEX-5.5	Enable audit logging	HIGH
VERTEX-5.6	Obtain HIPAA BAA	HIGH

New Files to Create (12 files, ~6,200 lines)

File	Lines	Description
vertex_ai_provider.py	~800	Core Vertex AI provider
vertex_config.py	~200	Configuration management
vertex_rag_engine.py	~600	RAG corpus management
vertex_vector_search.py	~500	Vector Search integration
vertex_agent_builder.py	~1,000	Agent Builder integration
vertex_mcp_tools.py	~600	MCP tool definitions
vertex_fine_tuning.py	~400	Model fine-tuning
vertex_security.py	~300	Security configuration

routers/vertex_rag.py	~400	RAG API endpoints
routers/vertex_search.py	~300	Search API endpoints
routers/vertex_agent.py	~500	Agent API endpoints
frontend/vertex_agent.js	~600	Agent chat UI

New API Endpoints (25+)

Vertex AI Provider

GET /api/vertex/status - Provider status

GET /api/vertex/models - Available models

POST /api/vertex/generate - Text generation

POST /api/vertex/embed - Generate embeddings

RAG Engine

POST /api/vertex/rag/corpus - Create corpus

GET /api/vertex/rag/corpus - List corpora

GET /api/vertex/rag/corpus/{id} - Get corpus details

DELETE /api/vertex/rag/corpus/{id} - Delete corpus

POST /api/vertex/rag/corpus/{id}/ingest - Ingest documents

POST /api/vertex/rag/corpus/{id}/query - Query with grounding

Vector Search

POST /api/vertex/search - Semantic search

POST /api/vertex/search/similar/{id} - Find similar competitors

GET /api/vertex/search/index/status - Index status

Agent Builder

POST /api/vertex/agent/session - Create session

POST /api/vertex/agent/chat - Send message

POST /api/vertex/agent/research/{id} - Research competitor

GET /api/vertex/agent/alerts - Get alerts

POST /api/vertex/agent/monitor/start - Start monitoring

Cost Analysis (~$78/month)

Service	Usage	Monthly Cost
Gemini 3 Flash (Input)	50M tokens	$7.50
Gemini 3 Flash (Output)	25M tokens	$15.00
Gemini 2.5 Pro (complex)	5M tokens	$6.25

Vector Search Queries	100K queries	$10.00
Vector Search Storage	10GB	$2.50
RAG Engine	10 corpora	Included
Agent Sessions	1,000 sessions	$20.00
Fine-Tuning (quarterly)	1 job	$16.67
TOTAL		**$77.92**

Success Metrics

Metric	Current	Target
AI response accuracy	~85%	>95%
Hallucination rate	~15%	<5%
Research time	30min manual	5min automated
News monitoring	Manual daily	Real-time automated
Battlecard freshness	Weekly manual	Auto-updated
Enterprise compliance	Not compliant	HIPAA compliant
Search relevance	Keyword only	Semantic

Risks and Mitigations

Risk	Impact	Mitigation
GCP service outage	High	Maintain Google AI SDK fallback
Cost overruns	Medium	Set budget alerts, optimize queries
Fine-tuning quality	Medium	Iterative training, human review
Agent hallucinations	Medium	Grounding required for all responses
Migration complexity	Medium	Phased rollout, feature flags

Conclusion

Integrating Vertex AI into Certify Intel will transform the platform from a manual competitive intelligence tool into an autonomous, enterprise-grade intelligence system. The ~$78/month investment delivers RAG-grounded responses, autonomous agents, semantic search, custom model fine-tuning, and HIPAA compliance - capabilities that would cost significantly more to build manually.

Recommended Next Step: Approve plan and begin Phase 1 with GCP project setup.