Architecture & Agent Design Report

Google Cloud NLP & Agentic Workflow Project

Prepared by: Joris Jose

Date: 14 September 2025

1. Technical Approach & GCP Services Used

The project addresses the challenge of extracting actionable insights from unstructured text such as reviews, tickets, and reports. The technical approach was guided by scalability, modularity, and leveraging Google Cloud's managed services for enterprise-grade reliability.

Service	Purpose	How It Is Applied
Google Cloud Natural Language API	Entity & Sentiment Extraction	Identifies names, dates, locations, and measures sentiment polarity & magnitude.
Vertex AI Generative Models	Summarization & Insight Generation	Produces concise summaries, abstracts, and knowledge synthesis for decision-making.
Cloud Storage	Data Lake	Stores raw inputs (CSV, JSON) and persists processed outputs for further analytics.
Python Agent Workflow	Orchestration	Coordinates the sequence of tasks—data preparation, extraction, summarization—with error handling.

2. Results, Challenges, and Trade-offs

Results:

- Delivered accurate entity recognition and sentiment scoring across diverse datasets. - Produced concise, human-readable summaries that improved downstream reporting and analytics. - Enabled structured insights suitable for dashboards, anomaly detection, and business intelligence.

Challenge	Description
Quota Management	Cloud NLP API limits required batching, queuing, and retry logic.
Input Size Constraints	Large documents had to be segmented without losing semantic context.
Latency vs. Accuracy	Balancing real-time needs against the depth of generative summarization.

Trade-offs:

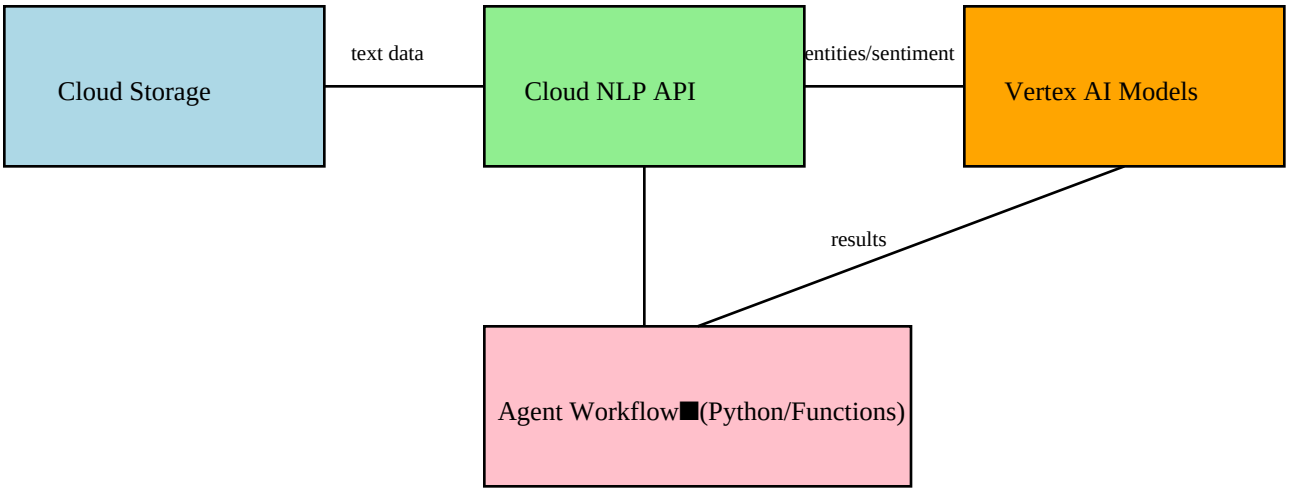
- **Cloud NLP API**: Fast, cost-effective, best for large-scale entity & sentiment tasks. - **Vertex AI**: Advanced generative summaries but higher latency and cost. - **Hybrid Strategy**: Applied each where most effective (NLP for extraction, Vertex AI for summarization).

3. Agentic Workflow & Architecture

The agentic workflow was designed as a modular, fault-tolerant system where each agent specializes in a task. This ensures maintainability, independent scaling, and extensibility. New agents (e.g., translation or classification) can be plugged into the pipeline with minimal changes.

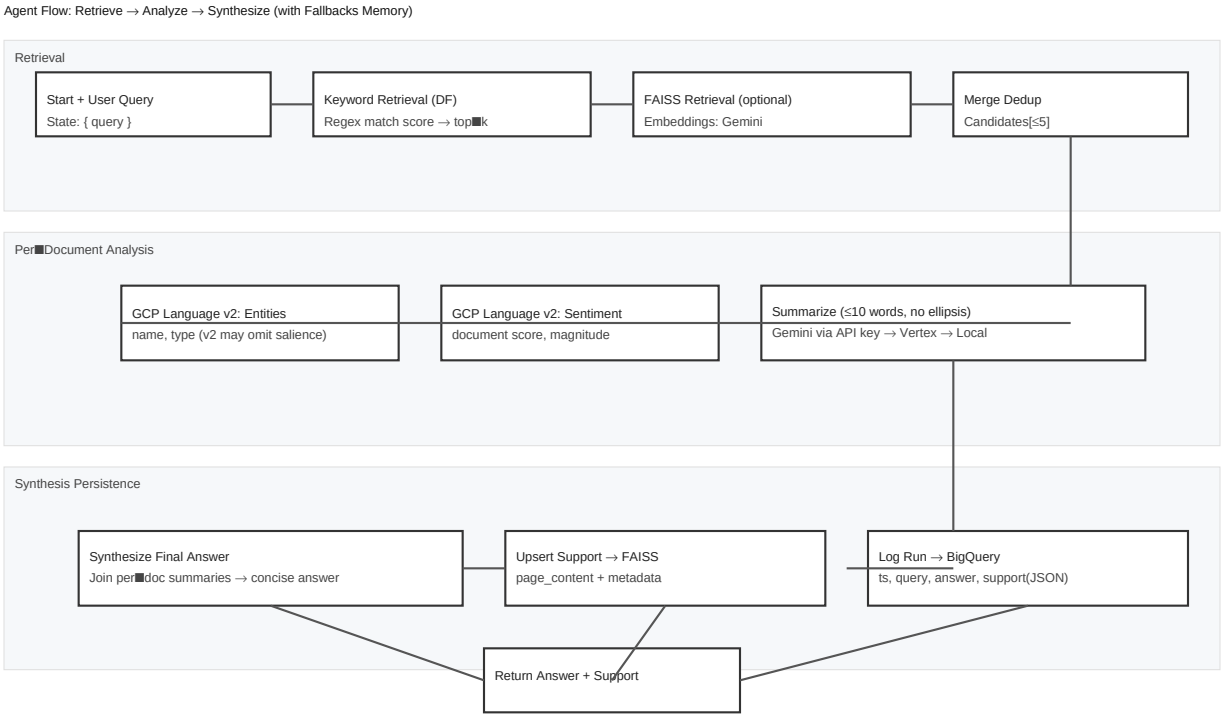
Stage	Description
Data Preparation	Raw data is cleaned, tokenized, normalized, and prepared for NLP.
Entity & Sentiment Extraction	Cloud NLP extracts entities, categories, and sentiment at document/row level.
Summarization & Knowledge Synthesis	Vertex AI generates concise summaries, highlighting critical information.
Orchestration Layer	Python workflow coordinates agents, manages retries, and ensures resilience against API failures.

4. High-level GCP Architecture



5. Agent Flow Diagram

This diagram visualizes the sequence of retrieval, tool calls (Cloud NLP, Vertex AI/Gemini), and synthesis steps. It highlights resilience mechanisms such as error handling and fallback logic.



Errors are captured per step; agent continues with available signals (graceful degradation).