Architecture & Agent Design (GCP First)

Overview
This prototype ingests unstructured text, extracts key information, and generates short summaries using managed GCP services.
I kept a local mode for quick iteration and a clean path to production on Vertex AI.

GCP Services
- Cloud Storage for raw input files
- BigQuery for structured corpus, extractions, and summaries
- Natural Language API for entities and sentiment
- Vertex AI for summarization
- Optional: Agent framework for planner and tool orchestration
- Cloud Logging and Monitoring for metrics and alerts
- Cloud Run or Cloud Functions for tool endpoints
- Pub/Sub and Vertex AI Pipelines for orchestration

High-Level Flow
Users or batch jobs write documents to Cloud Storage. An ingest job registers the documents in BigQuery.
Preprocess jobs normalize text. The Natural Language API extracts entities and sentiment. Vertex AI generates summaries.
Agent tools search and summarize documents. A simple planner calls these tools in sequence.

Agent Scenario
Question: What are the top issues and sentiment trends in recent product reports this week?
Plan: Search the corpus for the time window. Extract entities and issues. Summarize each relevant document. Aggregate results.

Trade-offs
Accuracy: Typed entities from the Natural Language API are reliable. Summaries are strong but require validation for critical use.
Latency: Batch processing reduces end-to-end latency and cost. Cache repeated summaries.
Cost: Use quotas, timeouts, and sampling. Partition BigQuery tables by date.

Productionization
- Scalability: Orchestrate with Vertex AI Pipelines and trigger with Pub/Sub
- Security and Privacy: IAM scoping, encryption, and prompt hygiene; redact PII in logs
- Monitoring: Collect latency, error, and tool success metrics; set SLOs and alerts
- CI/CD: Build and deploy with Cloud Build or GitHub Actions; use canary releases
- Governance: Label and document datasets and prompts; maintain a model registry
- Cost Controls: Budgets and alerts; autoscaling with a minimum of zero; BigQuery partitioning

I wrote and tested this in local mode, then outlined a clear path to harden and deploy on GCP.