Architecture & Agentic System Design Report

1. Overview & Technical Approach

This project implements an **agentic workflow** for intelligent news analytics using Google Cloud Platform (GCP) services. The system automatically extracts key entities, sentiment, and concise summaries from large-scale news datasets, enabling business insights, research summaries, or content monitoring.

GCP Services Used:

- **Cloud Storage:** Data lake for raw news records.
- **BigQuery (optional):** Scalable aggregation, analytics, and storage.
- **Natural Language API:** Entity extraction and sentiment analysis.
- **Vertex AI Gemini:** Advanced, generative summarization.
- **Vertex AI Pipelines (productionization):** Orchestration of workflows.
- **Cloud Functions, Pub/Sub (productionization):** Serverless triggers and messaging.
- **Firestore/Memorystore (optional):** Session memory and context persistence.

---

2. Scenario: Research Assistant Agent

Agent's Goal:

Automatically answer complex user queries about trends, sentiments, and key facts in news data, e.g.,
"Summarize the main topics and public sentiment regarding climate change in the last month."

Business Problem Solved:

- Rapid insights from massive, fast-changing news sources.
- Automated market, competitor, or risk monitoring.
- Efficient research for analysts or journalists.

---

3. Agentic Workflow & Reasoning

Workflow Steps

1. **User Query:**
   User requests a report ("Show me trending entities in tech news and summarize their sentiment.").

2. **Data Selection:**
   Agent selects relevant articles (by time, category, keywords) from Cloud Storage (or BigQuery).

3. **For Each Article:**
   - **Entity Extraction:** Use Natural Language API to find named entities (people, orgs, locations, etc.).
   - **Sentiment Analysis:** Use Natural Language API to score sentiment.

- **Summarization:** Use Vertex AI Gemini to generate one-sentence summary.

4. **Aggregation:**

   Group and count entities, aggregate sentiment, and collect summaries.

5. **Reasoning & Filtering:**

   Filter for top entities, unusual sentiment shifts, or requested trends.

6. **Report Generation:**

   Compose final report (table, chart, or natural language summary) and present to user.

7. **Ambiguity Handling:**

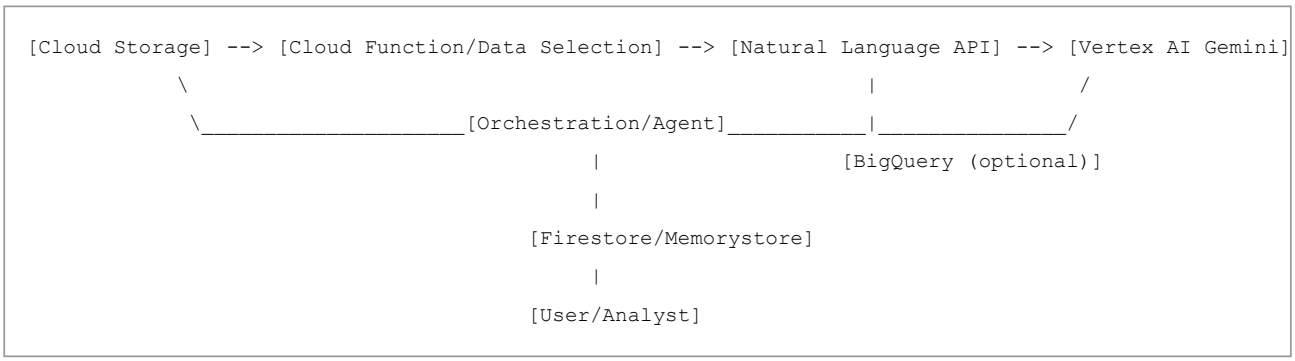
   If the query is unclear or results are insufficient, prompt the user for clarification or refine search.

8. **Memory:**

   Persist previous user queries and agent context in Firestore or BigQuery for personalized follow-up.

Flow Diagram

High-Level GCP Architecture Diagram



---

4. Results, Challenges & Trade-Offs

Results:

- High-quality extraction and summaries for most news articles.
- Efficient scaling and parallel processing via GCP APIs.

Challenges:

- Rate limits and quotas on NLP/GenAI APIs.
- Handling ambiguous or poorly structured news texts.
- Latency for large batch jobs.

Trade-Offs:

- GCP APIs offer high accuracy and ease of use, but can be costly for large volumes.
- Custom models could be tuned for domain but require more engineering.

---

5. Productionization Approach

To deploy this pipeline in production:

Scalability & Orchestration

- Use **Vertex AI Pipelines** for workflow orchestration and batch processing.
- Implement **Cloud Functions** for event-driven triggers (e.g., new data arrival).
- Use **Pub/Sub** for decoupled, scalable messaging between components.

Security & Data Privacy

- Enforce **IAM roles** for API and data access.
- Use **encryption** at rest (Cloud Storage, BigQuery) and in transit.
- Mask or anonymize sensitive data in outputs.

Monitoring & Logging

- Integrate **Cloud Logging** and **Cloud Monitoring** for tracking usage, errors, and performance.
- Set up automated alerts on failures or cost overages.

Cost Management & Optimization

- Use batching, caching, and only process relevant data to minimize API usage.
- Monitor billing and optimize through quotas and resource selection.

CI/CD & Reproducibility

- Use **Cloud Build** and GitHub Actions for automated testing and deployment.
- Version control all code and config.
- Containerize components for repeatable deployment.

---

6. Memory

For enhanced agent intelligence, persist user context and previous queries using **Firestore** or **BigQuery**.

This enables personalized recommendations, session recall, and longitudinal trend analysis.

---

7. Conclusion

This agentic GCP workflow supports scalable, robust news analytics with modular, cloud-native NLP and GenAI tools. The architecture and design allow for rapid prototyping, easy extension, and enterprise-grade productionization.