Reference Architecture Context:

Core Architecture Summary: AI-Powered Consumer Insights Application on Cloud Run with Hybrid Data Storage

This architecture powers a scalable, serverless, AI-driven Consumer Sense application deployed on Google Cloud Run. It processes consumer signals such as product reviews, ratings, pricing trends, competitor information, and multimodal inputs (files/web data) using agents, LLMs, and hybrid storage systems.

---

1. Application Layer

- **Application:**
  Handles core business logic, user queries, UI interactions, and insight generation flows.

- **Authentication (Auth):**
  Secures access to the application using OAuth/Firebase Auth.

---

2. Core Integration & Orchestration (MCP Toolbox & ADK)

- **MCP Toolbox:**
  Middleware/data abstraction layer that provides a unified interface for connecting to multiple storage layers (BigQuery, MongoDB Atlas, Cloud Storage, SQL, etc.).

- **ADK (Agent Development Kit):**
  AI agent framework responsible for orchestrating tasks like review analysis, category-level summarization, RAG retrieval, embeddings refresh, pricing intelligence, and NLP processing.

---

3. Data & Storage Layer (Hybrid Approach)

Implements polyglot persistence to support analytical, transactional, and unstructured data needs.

- **MongoDB Atlas:**
  NoSQL document store for reviews, enriched consumer signals, embeddings used for RAG/vector search.

- **BigQuery (BQ):**
  Serverless data warehouse for consumer trend analysis, pricing insights, competitor aggregates, and category models.

- **Cloud SQL (Optional):**
  Relational store for user metadata, sessions, and application configurations.

- **Cloud Storage:**
  Stores multimodal data such as PDFs, screenshots, scraped pages, product catalogs, and batch dumps.

---

4. AI & Context Layer

- **LLM (Large Language Model):**
  Powers AI reasoning and natural language understanding/generation for consumer insights. (Gemini)

- **Context Guardrails:**
  Ensures safe and domain-aligned AI outputs. Validates sentiment ranges, hallucination checks, and contextual consistency.

- **Web Data / Files Multimodal:**
  Inputs from websites, product listings, image files, and document uploads.
  Supports RAG pipelines and contextual enrichment.

---

5. Processing & Operations

- **Jobs:**
  Background tasks for scheduled scraping, embeddings refresh, sentiment recalculation, competitor sync, and batch processing.

Triggered via Cloud Run Jobs / Cloud Scheduler.

---

6. Development & AI Platform

- **Gemini (Implicit):**
  Underlying LLM used across the application for insight generation.

- **Firebase Studio:**
  Used for frontend/web app scaffolding where needed.

- **Gemini Code Assist:**
  Accelerates development and integration tasks.

Overall Flow:

The Application interacts through the MCP Toolbox and ADK to access multiple data sources.
The system leverages Gemini for AI capabilities, using multimodal data within Guardrails.
All core components run within a scalable Cloud Run environment with a hybrid storage backend.