**GovSight Core Storage Strategy: Local Archive + Pinecone + Web Retrieval Integration**

**🔍 Vision Statement:**

GovSight is being developed as a persistent, intelligent, AI-powered lobbying assistant that can track, analyze, and report on federal legislation, municipal priorities, and public sector issues. To function correctly, it must have a three-tier storage and retrieval system that supports semantic recall, deep legislative analysis, and real-time discovery.

**✅ Core Design Principle:**

**GovSight must use local storage, Pinecone vector search, and web retrieval as complementary layers.**

**💾 Local Storage (File System / Structured DB)**

**Purpose:** Archive and processing base

**Why it's needed:**

* To store *every version* of a bill, amendment, or meeting record over time
* To enable **diffing**, redlining, and temporal comparisons ("What changed this week?")
* To support structured queries on legislative actions, timestamps, roll calls, and sponsors
* To maintain a **source of truth** for rebuilding context, even if Pinecone is wiped or limited

**Example Use Cases:**

* Comparing bill texts from July 17 and July 24
* Reviewing all amendments to H.R. 1234
* Tracking actions taken in a specific committee or chamber

**Storage Format:**

* XML, JSON, Markdown, or SQLite
* Folder structure: /congress/118/hr/1234/2025-07-24.xml

**📈 Pinecone Vector Storage (Semantic Indexing)**

**Purpose:** Fast retrieval + RAG (Retrieval-Augmented Generation)

**Why it's needed:**

* Enables **semantic search** ("What bills relate to clean water in California?")
* Rapid **context injection** for LLM queries
* Stores **summaries, embeddings, chunked sections** of bills or meeting minutes
* Allows similarity searches and fast memory recall

**Example Use Cases:**

* Ask: "What similar projects have received funding in past 3 years?"
* Ask: "Summarize all amendments to this bill that affect the EPA"

**Embedding Target Content:**

* CRS summaries
* Committee reports
* Meeting minutes
* Parsed actions, sponsors, policy tags

**🌐 Web Retrieval Layer (Live Search)**

**Purpose:** Last-mile real-time access

**Why it's needed:**

* To handle queries when a fact or document is *not yet ingested locally* or in Pinecone
* Allows GovSight to stay up-to-date with breaking changes (e.g. new bills, recently updated websites)
* Ensures continuity and accuracy even if data is incomplete locally

**How it works:**

* When Pinecone and local return no confident answer → perform web search
* Scrape full page content (not just snippets)
* Parse and synthesize relevant facts or passages
* Optionally ingest and persist relevant content for future use

**Example Use Cases:**

* “Who is the new mayor of Grandview, TX?”
* “What committee hearing is scheduled for H.R. 4567 this week?”
* “Did Rep. Garcia introduce a new bill on water treatment?”

**⚖️ Why All Three Are Needed**

| **Functionality** | **Local Storage** | **Pinecone** | **Web Retrieval** |
| --- | --- | --- | --- |
| Full-text versioning and diffing | ✅ Yes | ❌ No | ❌ No |
| Fast semantic search (e.g., similar bills) | ❌ No | ✅ Yes | ❌ No |
| Audit trail and compliance | ✅ Yes | ❌ No | ❌ No |
| Source of truth | ✅ Yes | ❌ No | ❌ No |
| Embedding-based querying | ❌ No | ✅ Yes | ❌ No |
| Real-time fact lookups | ❌ No | ❌ No | ✅ Yes |
| Fallback for missing info | ❌ No | ❌ No | ✅ Yes |
| Offline resilience | ✅ Yes | ❌ No | ❌ No |

**✅ Final Design Mandate:**

**All core functionality in GovSight (e.g. "what changed this week?", "who sponsored this amendment?", "what’s similar to X?", "what’s the current mayor’s name?") depends on a layered system that integrates a local archive, a semantically searchable Pinecone index, and real-time web retrieval.**

These three layers must be kept in sync through ingestion logic, fallback chaining, and timestamped records to ensure accuracy, responsiveness, and completeness over time.

**✨ Future Reference:**

This file summarizes the three-tier GovSight storage and retrieval model and must be reloaded in every future dev session if lost. This is **not optional**, it is **central to how GovSight works.**