

Cuestión de Datos

Copiloto de IA para el Diseño de Políticas Públicas Basadas en Evidencia

Juan Camilo Grajales - DG Capital
Diciembre 2025
Concurso: Datos al Ecosistema - MinTIC

1. Resumen Ejecutivo

En Colombia, la formulación de políticas públicas enfrenta una brecha crítica entre la intención política y la evidencia técnica. Los datos necesarios para diseñar intervenciones efectivas (pobreza, salud, educación, inversión) se encuentran dispersos en silos de información gubernamentales (DNP, MinSalud, APC, etc.), con barreras técnicas y semánticas que dificultan su integración en tiempo real por parte de los funcionarios.

Cuestión de Datos es una plataforma web que introduce un nuevo paradigma: el "Policy Canvas" asistido por un Agente de IA. A través de una arquitectura RAG Agéntica (Retrieval-Augmented Generation) basada en el patrón ReAct (Reason + Act), la herramienta permite a los usuarios redactar documentos de política pública mientras un copiloto de inteligencia artificial "lee" el borrador, identifica conceptos clave y entidades territoriales, y ejecuta consultas SQL (SoQL) autónomas sobre las APIs de Datos Abiertos del Estado (Socrata).

La solución no solo recupera datos, sino que resuelve problemas complejos de interoperabilidad semántica mediante un middleware de normalización geográfica propio (DIVIPOLA), ofreciendo funciones avanzadas como la auditoría de riesgos fiscales y la integración directa de la evidencia en la narrativa del documento final.

2. El Problema: Fragmentación y Ceguera Estratégica

El ciclo de vida de las políticas públicas en Colombia se ve obstaculizado por:

- **Silos de Información:** Los datos críticos residen en sistemas desconectados. Un formulador de proyectos no puede correlacionar fácilmente la demanda social (Sisbén IV, DNP) con la oferta institucional (Inversión APC) o los resultados de calidad (RIPS, MinSalud).
- **Barreras Técnicas:** El acceso a datos abiertos requiere conocimientos de consulta (SQL, APIs) que exceden el perfil promedio de un funcionario público.
- **Heterogeneidad Semántica:** La falta de estándares transversales hace que una misma entidad (ej: "Medellín") se represente de formas incompatibles entre bases de datos (Medellín, MEDELLIN, 05001, Colombia/Antioquia/Medellin), impidiendo cruces automáticos.
- **Riesgo de "Alucinación" en IA:** Los modelos de lenguaje tradicionales (LLMs) no son fiables para datos factuales, ya que tienden a inventar información cuando no tienen acceso a fuentes externas en tiempo real.

Esto resulta en una "ceguera estratégica", donde las decisiones se basan en intuición o datos obsoletos, aumentando el riesgo de ineficiencia fiscal y proyectos fallidos ("elefantes blancos").

3. La Solución: Arquitectura RAG Agéntica (ReAct)

Cuestión de Datos supera estas limitaciones implementando una arquitectura Agéntica sobre el modelo RAG (Retrieval-Augmented Generation), potenciada por el patrón de razonamiento ReAct.

3.1. Componentes Core

- Modelo de Lenguaje (LLM): Google Gemini (vía API REST).
- Backend: Next.js API Routes (Serverless Functions).
- Fuentes de Datos: APIs SODA (Socrata Open Data API) de 5 datasets estratégicos de Colombia (Salud, Educación, Sisbén, APC, Senado).
- Middleware de Normalización: Módulo propio (maestro_divipola.js) para estandarización geográfica.

3.2. Flujo de Razonamiento ReAct

A diferencia de un chatbot RAG estándar que hace una sola búsqueda y responde, el agente de Cuestión de Datos sigue un ciclo iterativo:

- Pensamiento (Reason): El agente analiza el contexto del usuario (ej: "Redactando sobre salud en Medellín").
- Acción (Act): Decide qué herramienta usar.
- Middleware: Si detecta una entidad geográfica, invoca al middleware de normalización.
- Exploración: Si necesita explorar datos, genera una consulta SoQL (SELECT ... WHERE ... LIKE ...) y la ejecuta contra Socrata.
- Observación (Observe): Recibe la respuesta cruda de la API (JSON de datos o error).
- Repetición/Respuesta: Si la observación es suficiente, genera la respuesta final. Si no, refina su pensamiento y ejecuta una nueva acción (hasta 5 intentos).

3.3. Middleware de Normalización Geográfica (El "Pegamento")

Para resolver la heterogeneidad semántica, se desarrolló un sistema de resolución de entidades basado en el estándar DIVIPOLA del DANE. Este módulo intercepta las intenciones de búsqueda geográfica y las traduce al "dialecto" específico de cada dataset objetivo mediante el uso de comodines SQL estratégicos (ej: transformar "Medellín" en LIKE '%MEDELL_N%' para cubrir variantes con y sin tilde).

4. Funcionalidades Clave y Experiencia de Usuario (UX)

La interfaz se basa en un "Policy Canvas" de pantalla dividida, diseñado para el flujo de trabajo real del funcionario:

- Lienzo de Redacción (Izquierda): Un editor estructurado donde el usuario redacta su documento, con soporte para plantillas oficiales (MGA, CONPES, Policy Brief).
- Copiloto de Datos (Derecha): Un panel lateral donde el agente ReAct opera.
- Interacción "Investigar": Al hacer clic en "Investigar" sobre un párrafo, el sistema envía el contexto al agente, que inicia su ciclo de búsqueda autónoma.
- Integración Narrativa: Los resultados no se muestran solo como tablas crudas. El agente genera un párrafo técnico (narrativa basada en evidencia) que el usuario puede insertar directamente en su documento con un clic ("Copiar al Lienzo").
- Auditoría de Riesgos: Una función especializada donde el agente busca activamente datos que contradigan la propuesta del usuario (ej: buscar proyectos similares no ejecutados o techos presupuestales), actuando como un "abogado del diablo" fiscal.

5. Impacto y Valor Público

El impacto de Cuestión de Datos se refleja en varios frentes de valor público:

- Eficiencia Fiscal: Reduce la incertidumbre en la fase de pre-inversión al alertar sobre riesgos y duplicidades antes de la asignación de recursos.
- Transparencia Activa: Habilita a veedurías y ciudadanos para auditar técnicamente las propuestas gubernamentales, democratizando el acceso a la evidencia.
- Interoperabilidad Semántica: Provee un modelo replicable y de código abierto para superar las barreras de integración de datos en el Estado sin requerir costosos proyectos de infraestructura de backend.
- Adopción Tecnológica: Reduce la barrera de entrada al uso de datos abiertos, permitiendo que funcionarios sin perfil técnico aprovechen el valor de la información pública.

6. Conclusión

Cuestión de Datos no es solo una herramienta de consulta; es una plataforma de gobernanza inteligente. Al unir la capacidad de razonamiento de la IA generativa con la verdad objetiva de los datos abiertos del Estado, transforma la formulación de políticas públicas de un ejercicio retórico a una disciplina basada en evidencia, marcando el camino hacia un Estado más eficiente y transparente.

Repositorio GitHub: <https://github.com/Grajales-Camilo/cuestion-de-datos-MinTIC-2025>

URL del Proyecto: cuestiondedatos.com

Licencia: MIT (Open Source)