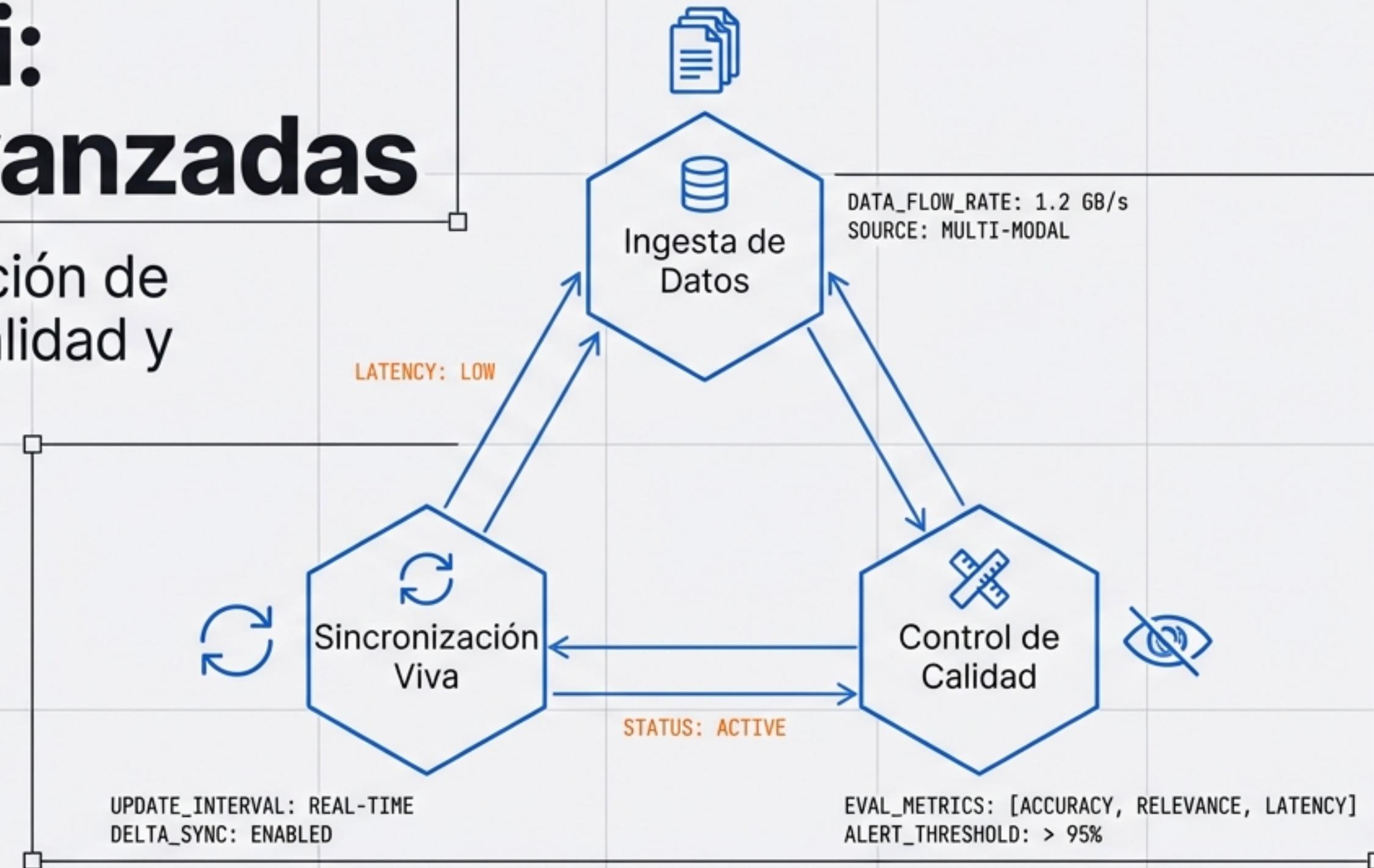


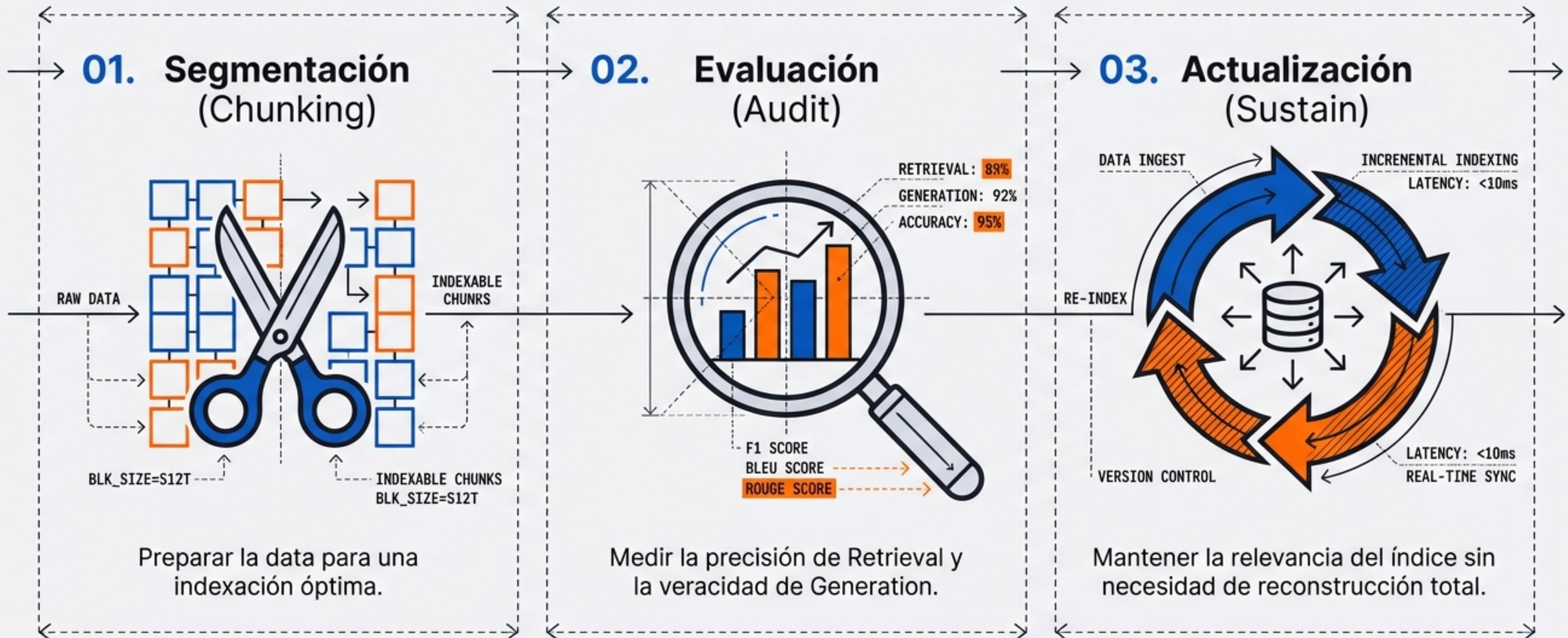
# RAG en Producción con Google Gemini: Estrategias Avanzadas

Guía técnica para Optimización de Chunking, Evaluación de Calidad y Actualización Incremental.



# El Ciclo de Vida de Producción RAG

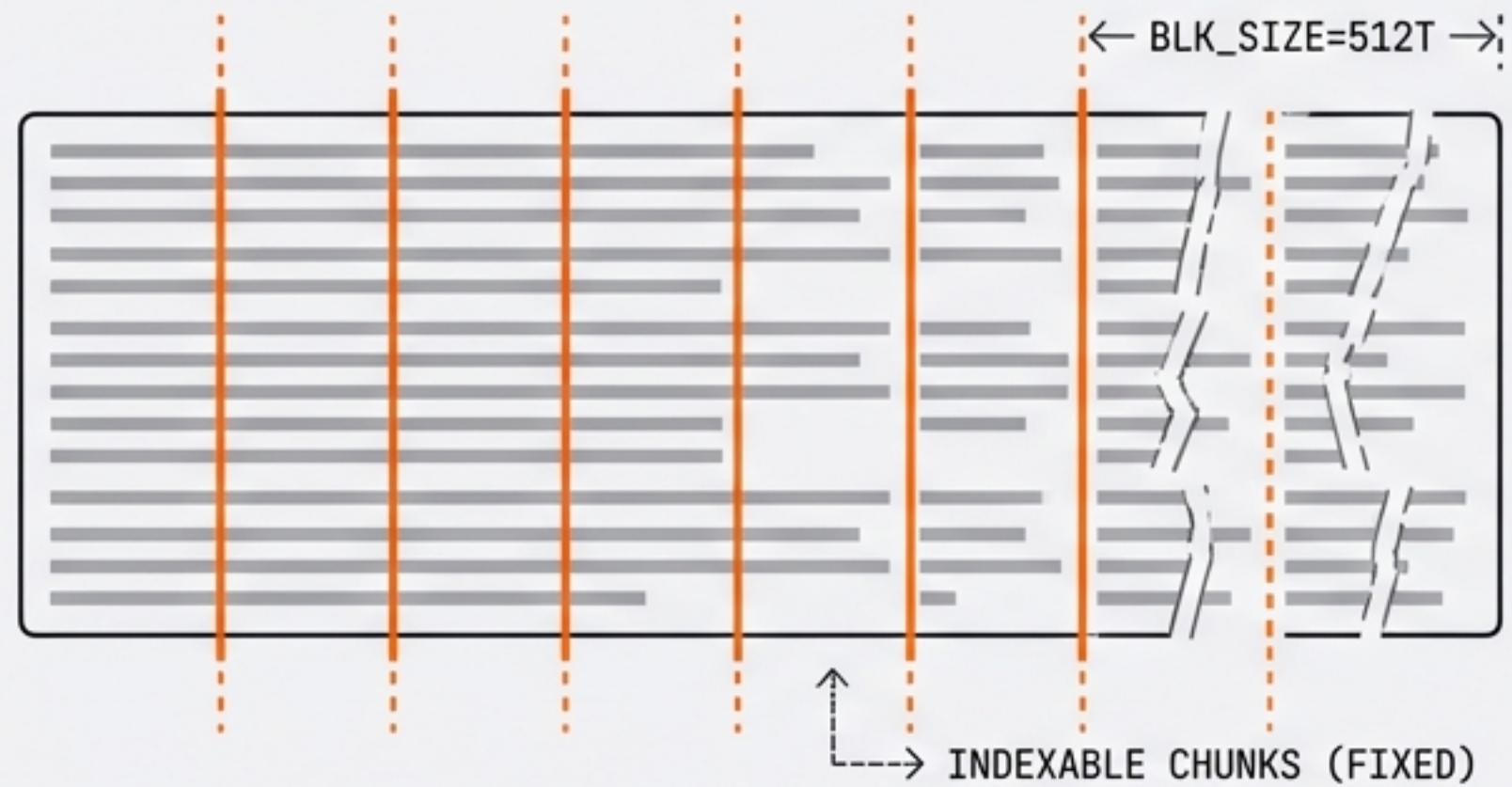
Un sistema RAG no termina en el despliegue; requiere una arquitectura de datos viva, medible y mantenible.



# Estrategias de Chunking: El Arte de la Segmentación

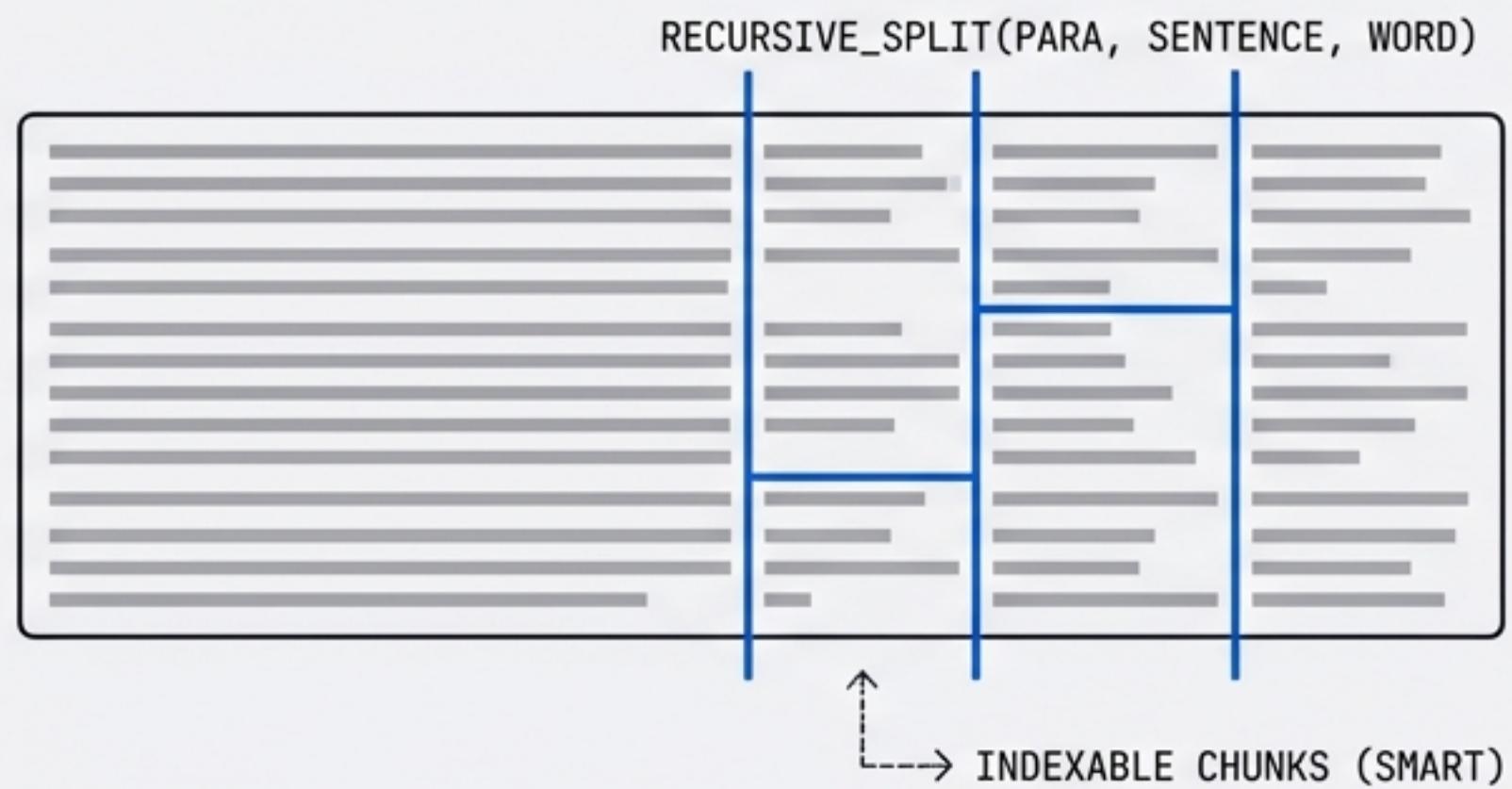
Métodos de baja complejidad para procesamiento rápido

## Método A: Fixed-Size Chunking (Tamaño Fijo)



- Ideal para texto homogéneo.  
Rápido, complejidad baja.

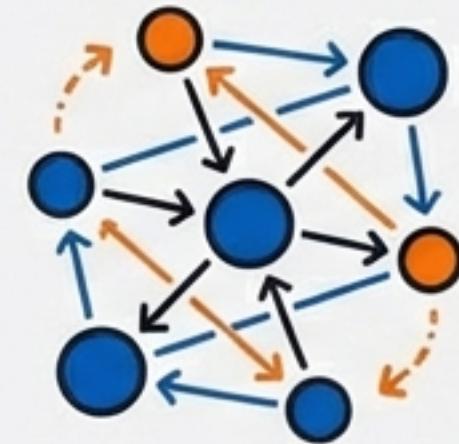
## Método B: Recursive Character Splitting



- División inteligente que respeta la gramática.  
Propósito general, complejidad media.

**Insight:** El chunking impacta directamente la calidad del RAG; fragmentos mal cortados rompen el contexto semántico.

# Chunking de Alta Precisión y Contexto



## Semantic Chunking

Agrupa texto basado en significado y similitud de embeddings.

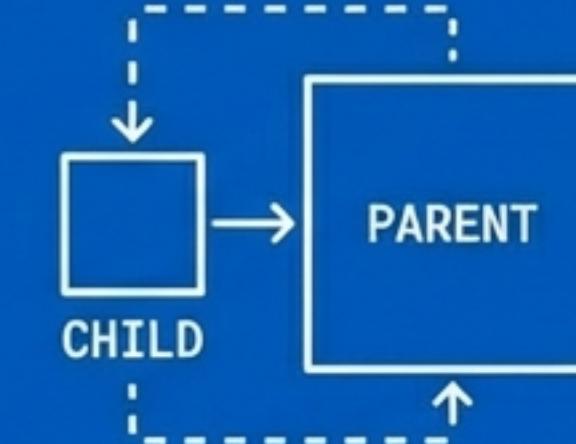
**Mejor para:** Documentos largos y complejos.



## Document-Structure Chunking

Respeta encabezados de Markdown o HTML.

**Mejor para:** Archivos estructurados.



## Parent-Child Chunking

Vincula fragmentos pequeños (para búsqueda precisa) con fragmentos padres más grandes (para contexto rico al generar).

! El Parent-Child Chunking ofrece el mejor balance entre búsqueda precisa y contexto extendido, ideal para RAG de alta fidelidad.

# Matriz de Selección de Estrategia

Estrategia	Mejor Para	Complejidad
Fixed-Size	Texto homogéneo	Baja
Recursive	Documentos generales	Media
Semantic	Documentos largos	Alta
Structure	Markdown, HTML	Media
Parent-Child	RAG de precisión	Alta
Code Chunking	Código fuente	Media (Requiere lógica basada en sintaxis)

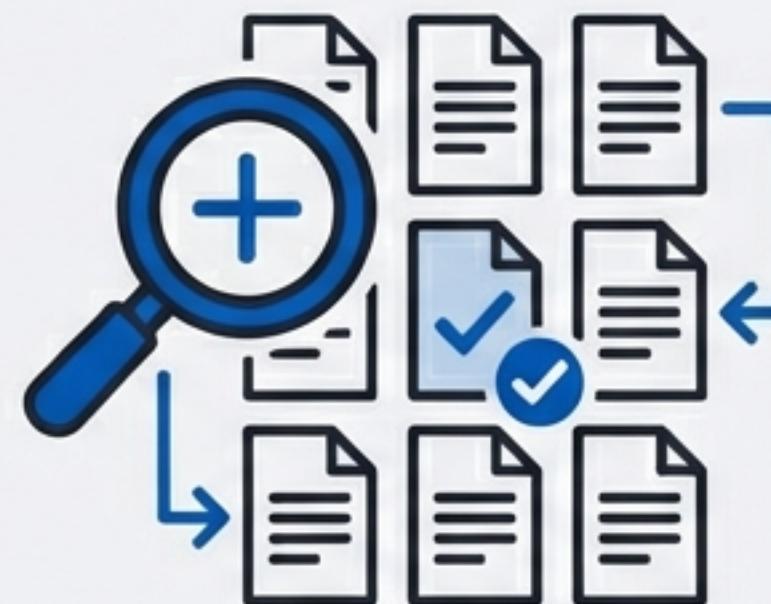
## Checklist de Configuración

- Tamaño típico: 200-500 tokens
- Incluir overlap (superposición) para continuidad
- Preservar metadata de origen

# Marco de Evaluación: Retrieval vs. Generación

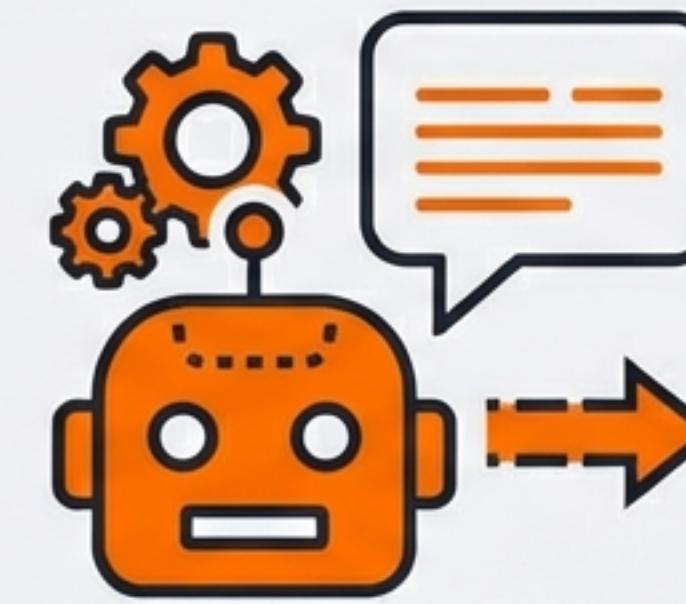
Dos puntos críticos de fallo en sistemas RAG.

## Retrieval (El Buscador)



¿Encontré los documentos correctos?

## Generation (El Redactor)



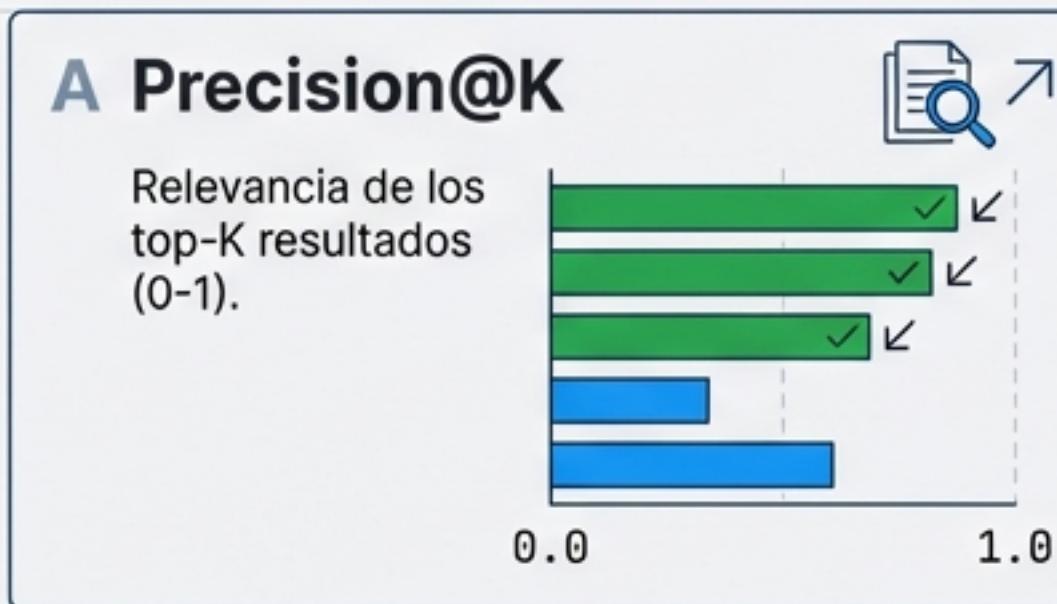
¿Es la respuesta útil y cierta?

**Requisito:** Dataset de Evaluación con 'Ground Truth' (la verdad terreno) para comparar las salidas.

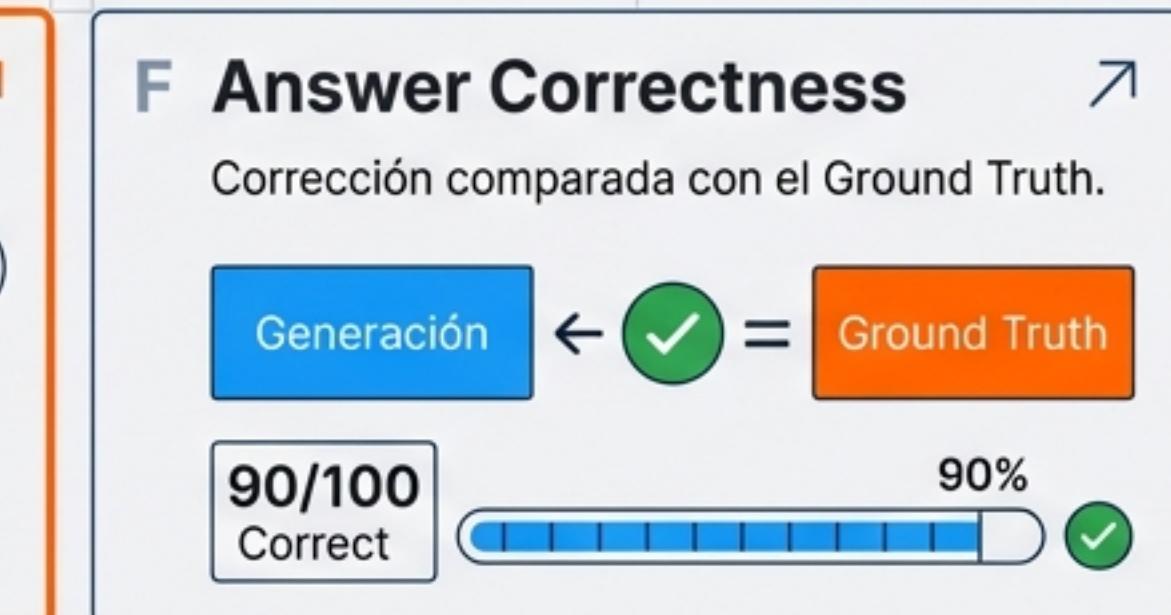
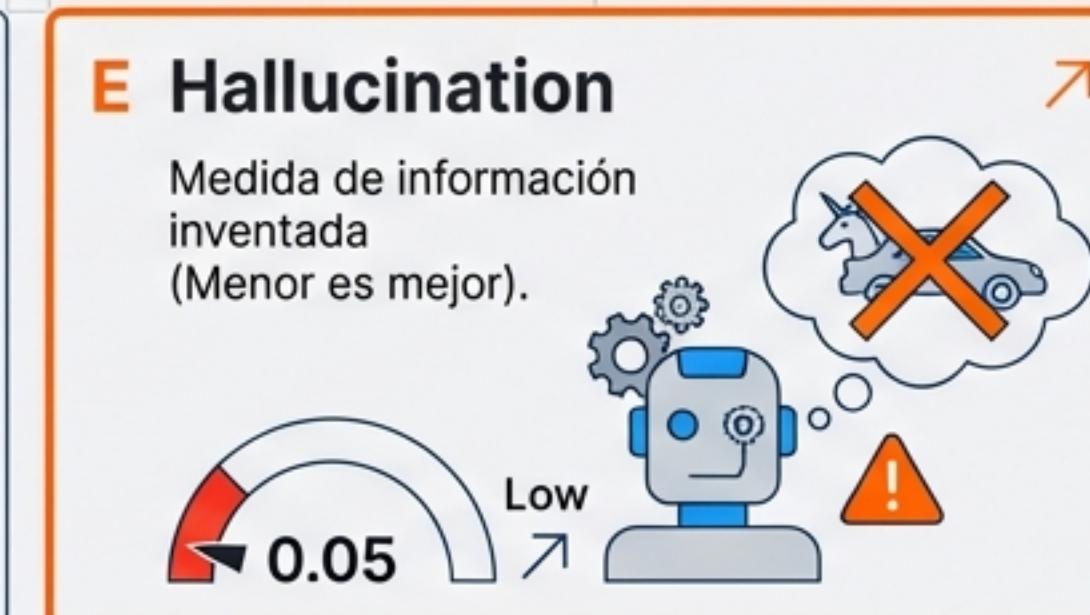
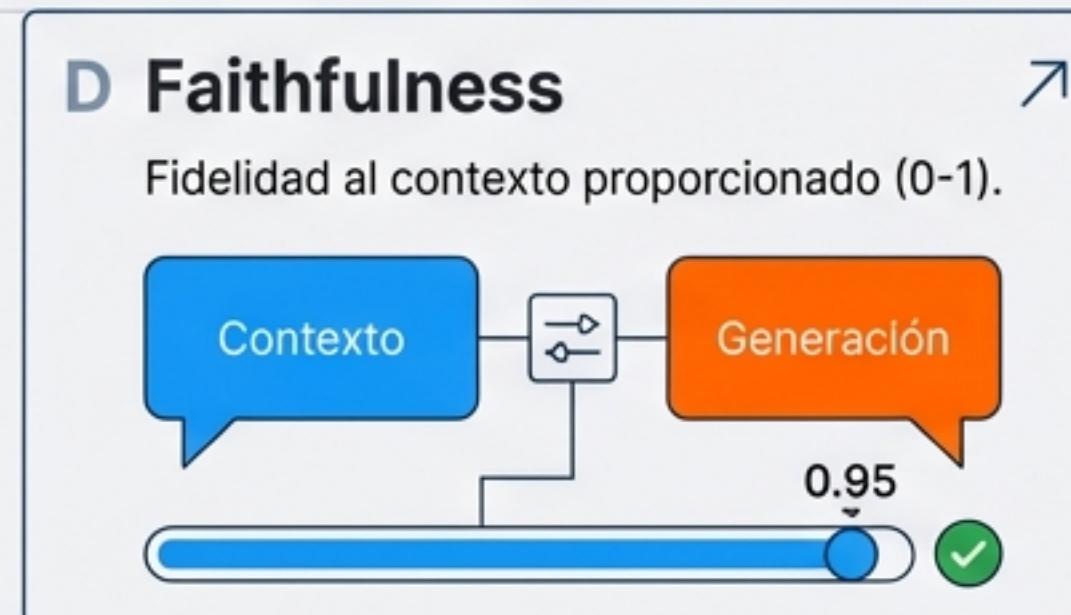
# Dashboard de Métricas Críticas

Transición de prototipo a ingeniería de sistemas

## Retrieval Metrics



## Generation Metrics



# Automatización con LLM-as-Judge

Usando Gemini para evaluar la calidad a escala.



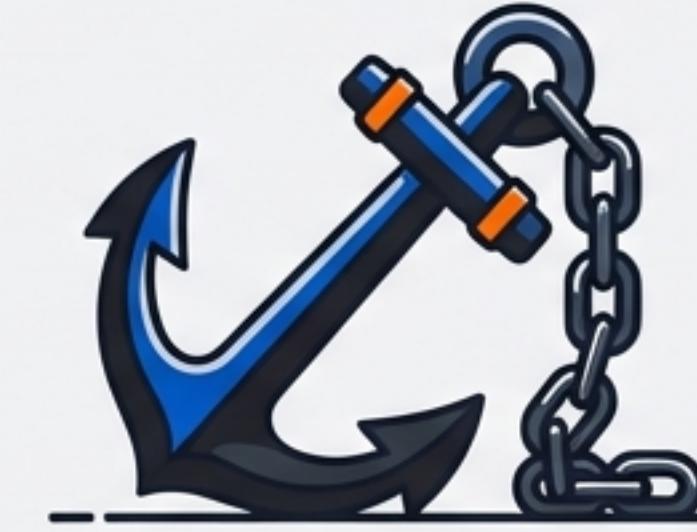
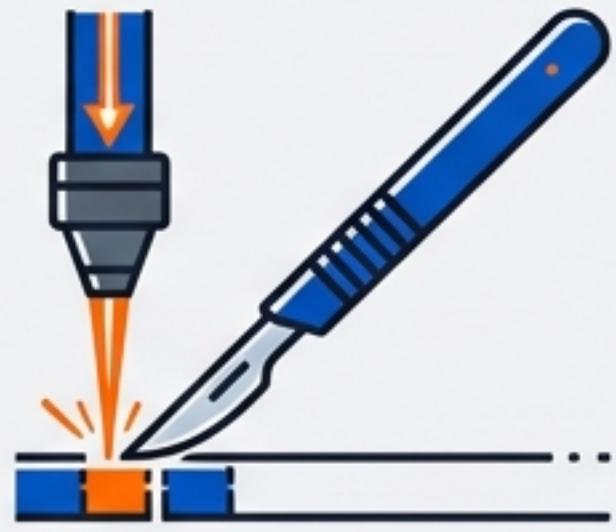
## Checklist de Implementación

- Crear dataset con ground truth.
- Configurar evaluación de recuperación y generación.
- Establecer umbrales de calidad aceptable.

# El Desafío de la Frescura: Actualización Incremental

Problem Statement

Data Drift: En producción, los datos cambian constantemente. Un índice estático se vuelve obsoleto rápidamente.

Re-indexación Total	Actualización Incremental
	

Lento, costoso, reconstruye todo desde cero.

Rápido, preciso, solo modifica lo que cambió (Aregar/Modificar/Eliminar).

# Arquitectura de Sincronización

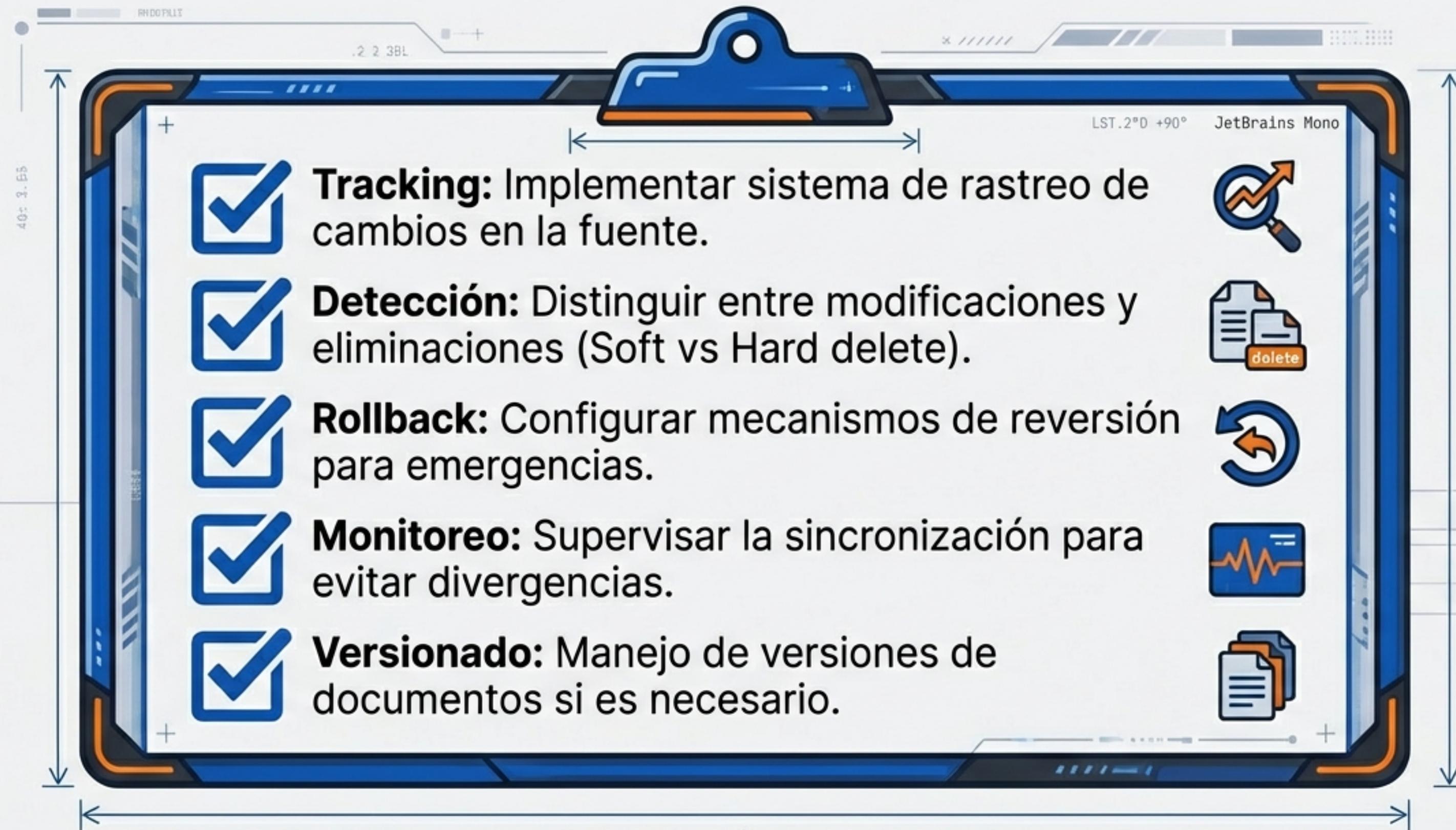


## Estrategias de Actualización

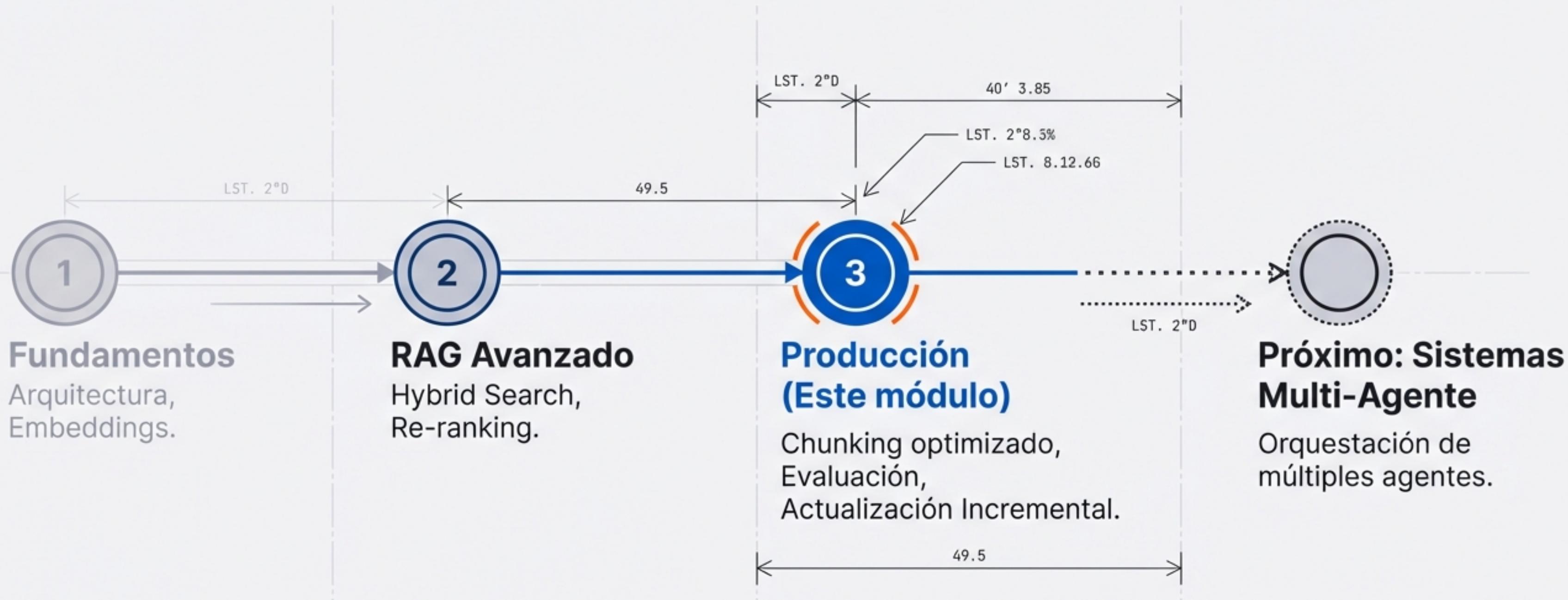
	Estrategia	Latencia	Complejidad	Caso
1	Tiempo Real	<1 min	Alta	Apps interactivas
2	Programada (Batch)	Minutos-Horas	Media	Procesamiento por lotes
3	Webhook	Segundos	Media	Sistemas integrados
4	Manual	Variable	Baja	Mantenimiento esporádico

30° 50' N 9° 50' E

# Checklist de Seguridad e Implementación



# Resumen del Módulo 6: RAG con Agentes



Dominar estos tres pilares asegura el paso exitoso del prototipo a la producción empresarial robusta.