

Estudio de Mercado

Soluciones de Resumen de Documentos con IA

Análisis comparativo de alternativas en la nube y on-premise

Preparado por: *Diana Morales Suárez*

Fecha: septiembre 4 de 2025

Contenido

1.	Resumen Ejecutivo	3
2.	Introducción.....	3
2.1	Contexto y análisis	4
2.2	Tamaño de la oportunidad de mercado.....	4
2.3	Objetivo del documento	4
2.4	Alcance y Limitaciones.....	5
3.	Metodología.....	6
3.1	Fuente de información consultadas	6
3.2	Criterios de comparación.....	6
4.	Análisis de Soluciones en la Nube.....	7
4.1	Descripción de herramientas revisadas	7
4.2	Cuadro Costos, Modalidad de Cobro y Límites.....	8
5.	Análisis de Soluciones On-Premise.....	10
5.1	Descripción de herramientas revisadas	10
5.2	Cuadro Costos, Modalidad de Cobro y Límites.....	11
6.	Comparativo (Nube vs. On-Premise).....	12
6.1.	Cuadro comparativo General	12
6.2.	Cuadro comparativo costos	13
6.3.	Cuadro comparativo Exportación de resúmenes	13
7.	Hallazgos clave.....	15
7.1.	Fortalezas del mercado actual	15
7.2.	Vacíos y brechas identificadas.....	15
8.	Oportunidad para dIAAna Insights	16
8.1.	Posicionamiento potencial	16
8.2.	Diferenciales estratégicos	16
9.	Conclusiones y Recomendaciones.....	18
9.1.	Síntesis del estudio	18
9.2.	Casos de uso concretos.....	18
9.3.	Lineamientos sugeridos para avanzar.....	18

1. Resumen Ejecutivo

El presente documento analiza el mercado de soluciones de Inteligencia Artificial enfocadas en la generación de resúmenes de documentos, particularmente en formatos como PDF. El estudio compara herramientas disponibles en la **nube** (ej. Adobe Acrobat AI, Microsoft Copilot, Google Gemini, Dropbox AI, entre otras) frente a alternativas **on-premise** y de **código abierto** (Ollama, privateGPT, llama.cpp, IBM Watsonx, NVIDIA NeMo, Haystack).

Los principales hallazgos son:

- **Modelos en la nube** ofrecen facilidad de uso, rápida adopción y costos predecibles (USD 10–30 mensuales por usuario, o cobro por uso de tokens). Sin embargo, presentan limitaciones de procesamiento (tokens, páginas o consultas), dependencia de proveedores externos y riesgos de seguridad en el manejo de datos sensibles.
- **Soluciones on-premise** destacan por no tener límites de páginas ni costos recurrentes de suscripción, pero requieren inversión en infraestructura tecnológica y conocimientos técnicos avanzados. En entornos Enterprise (ej. IBM, NVIDIA) los costos son elevados, aunque ofrecen control total sobre datos y flexibilidad de integración.
- Existe un **vacío de mercado** entre la simplicidad de las soluciones en la nube y el control de las on-premise.

En este contexto, **DIANA Insights** tiene la oportunidad de posicionarse como una solución **híbrida**, que combine:

- Facilidad de uso tipo nube.
- Capacidad de ejecución en infraestructuras locales del cliente.
- Exportación directa de resultados en PDF/Word.
- Procesamiento en español y posibilidad de personalización al contexto del negocio.

Con estos diferenciales, la compañía puede capitalizar una **ventaja competitiva clara** frente a competidores globales y open source, consolidándose como una alternativa confiable, flexible y adaptada al mercado regional.

2. Introducción

2.1 Contexto y análisis

Este documento presenta un estudio de mercado sobre las principales soluciones de Inteligencia Artificial enfocadas en la generación de resúmenes a partir de documentos (PDF y otros formatos). El análisis compara tanto herramientas disponibles en la nube como alternativas on-premise, con el objetivo de identificar ventajas, limitaciones y oportunidades de diferenciación

que pueden ser relevantes para la DB SYSTEM en su estrategia de innovación y desarrollo de producto.

2.2 Tamaño de la oportunidad de mercado

- **Global:**
 - El mercado de soluciones de *document AI* y *intelligent document processing* está en crecimiento acelerado. Según Gartner e IDC, se estima que supere los **USD 10.000 millones en 2027**, impulsado por sectores como legal, financiero y salud.
 - La necesidad no es solo resumir: es **entender, estructurar y extraer valor de los documentos**.
- **Latinoamérica:**
 - La adopción de IA va en aumento, pero con una brecha clara: la mayoría de soluciones internacionales **no están optimizadas para español** ni para contextos donde la infraestructura no siempre es de última generación.
 - Esto deja un espacio para una solución que se adapte a realidades locales.
- **Oportunidad para DB SYSTEM:**
 - Clientes potenciales: **gobierno, banca, universidades, sector legal y asegurador**, que procesan miles de PDFs y necesitan insights rápidos.
 - Si DB SYSTEM posiciona *DIANA Insights* en uno o dos de estos sectores primero, puede crecer como caso de éxito regional.

2.3 Objetivo del documento

- Identificar las principales soluciones de mercado en la nube y on-premise para resúmenes de documentos.
- Analizar costos, modalidad de cobro y límites de procesamiento.
- Evaluar características diferenciales (exportación de resultados, facilidad de integración, seguridad).
- Proponer conclusiones sobre posibles oportunidades para DIANA Insights en este contexto.

2.4 Alcance y Limitaciones

El estudio se centra en soluciones reconocidas a nivel global y en opciones de código abierto utilizadas en entornos empresariales. No se incluyen herramientas de nicho sin adopción significativa o con disponibilidad restringida a investigación.



Experiencia e innovación en proyectos TI

 Tel (601) 390 70 13

 www.db-system.com

 mercadeo@db-system.com

 Calle 97A N° 53 - 01 bogotá - Colombia

3. Metodología

El análisis se desarrolló a partir de la identificación de las principales soluciones tecnológicas en modalidad **nube** y **on-premise**, enfocadas en procesamiento, costos y escalabilidad. Para ello se realizó una revisión comparativa que consideró tanto variables técnicas como financieras, con el fin de ofrecer a la junta directiva una visión objetiva y práctica de los beneficios y limitaciones de cada enfoque.

3.1 Fuente de información consultadas

- Documentación oficial de fabricantes: fichas técnicas, manuales y whitepapers.
- Estudios de mercado y tendencias de la industria (Gartner, Forrester, IDC).
- Casos de uso y experiencias previas de empresas del sector.

3.2 Criterios de comparación

- **Costo de suscripción:** valor base de acceso a la solución.
- **Modalidad de cobro:** mensual, anual, por uso o licenciamiento fijo.
- **Límite de procesamiento de páginas:** restricciones en cantidad de documentos o tokens procesables.
- **Modalidad de despliegue:** cloud (software como servicio) u on-premise (infraestructura propia del cliente).
- **Tecnología detrás de la solución:** frameworks, motores y modelos de lenguaje empleados.

4. Análisis de Soluciones en la Nube

4.1 Descripción de herramientas revisadas

Producto / proveedor	¿Qué hace con PDFs?	Tecnología detrás (modelos/plataforma)	Seguridad / notas clave
Adobe Acrobat AI Assistant	Resumen, Q&A sobre PDF, genera puntos clave y citas	Modelos de Adobe + orquestación en Acrobat; asistente integrado en Reader/Acrobat	Datos no se usan para entrenar por defecto; controles empresariales.
Google Gemini en Drive / Vertex AI	En Drive/Workspace resume PDFs; en Vertex AI admite PDFs como input para resumir	Gemini (Vertex AI); APIs de <i>document understanding</i> y ejemplos de "Process a PDF"	Controles de Google Cloud/Workspace; opciones de residencia de datos en GCP.
Microsoft Copilot (Edge / 365)	En Edge : resumir PDF abierto; en M365: resumen y Q&A con contenido empresarial	Modelos de Azure OpenAI integrados; Copilot sobre Graph y OneDrive/SharePoint	Hereda cumplimiento de M365/Azure; controles IT.
Box AI	Resumen y Q&A sobre archivos (incluye PDFs) dentro de Box	Usa modelos de OpenAI y Anthropic (según Box)	Datos se quedan en Box; gobierno y permisos nativos.
Dropbox AI	Resumen y Q&A de archivos almacenados (PDF incluido)	Orquestación propia + LLMs (no detallan modelos específicos públicamente)	Integrado en Dropbox con controles empresariales.
ChatGPT (OpenAI)	Cargar PDF y pedir resumen/insights; Enterprise soporta visual retrieval en PDFs	OpenAI GPT-4.x/4.1 ; carga de archivos en ChatGPT/Enterprise	En Enterprise: aislamiento de datos y SSO; guías oficiales de archivos.
Perplexity (Pro)	Subes PDFs y te devuelve resúmenes y respuestas	Orquestación con LLMs + búsqueda	Foco en citas; plan Pro.
Humata	Sube PDFs; hace resúmenes, Q&A y extracción	LLMs + RAG sobre tus documentos	Enfatiza privacidad/seguridad para empresa.
SciSpace Copilot	Resumen de papers PDF, explicación sección por sección	LLMs ajustados a papers científicos	Muy usado para investigación académica.
AWS Bedrock (Knowledge Bases)	Construir apps que resuman/consulten PDFs vía RAG gestionado	Amazon Bedrock (Anthropic, Amazon, etc.); <i>Knowledge Bases</i> , GraphRAG	Servicio totalmente gestionado; integra parsing de documentos complejos.

4.2 Cuadro Costos, Modalidad de Cobro y Límites

Solución	Precio / Modalidad	Modalidad de Pago	Límite de Archivos o Páginas
Adobe Acrobat AI Assistant	\$4.99/mes (intro hasta junio 2025); luego \$9.99/mes	Suscripción mensual (add-on a Acrobat)	Hasta 10 documentos para "generative summary"
Google Gemini (Drive / Vertex AI)	API: pago por tokens (p. ej. Gemini 2.5 Pro: \$1.25 por millón input tokens, \$10 salida tokens) . También "AI Pro" y "AI Ultra" mensual (\$19.99 y \$249.99)	Pago por uso (tokens) y suscripciones mensuales	No hay límite de páginas; depende de tokens usados
Microsoft Copilot (365 / Edge)	Copilot incluido en M365; precio aprox. \$30–\$31.50/usuario/mes (<i>no encontramos fuente exacta, podrías consultarlo internamente</i>)	Suscripción mensual por usuario	No especificado públicamente
Box AI	Box Business desde \$20 a \$50/usuario/mes	Suscripción mensual por usuario	Tamaño por archivo entre 5 GB y 150 GB según plan
Dropbox AI	Dropbox Business \$20–\$26/usuario/mes	Suscripción mensual por usuario	Límite de tamaño según plan (hasta ~150 GB)
ChatGPT (OpenAI)	Plus \$20/mes; Team \$25–\$30/usuario; Pro \$200/mes; Enterprise (varía)	Suscripción mensual/anual por usuario	Carga de archivos hasta 512 MB (~2 M tokens)
Perplexity (Pro / Enterprise)	Pro \$20/mes o \$200/año; Enterprise Pro \$40/mes o \$400/año por usuario	Suscripción mensual/anual por usuario	Pro: uploads ilimitados; límite de consultas (~300/día)
Humata	Free: hasta 60 páginas; Student \$199/mes (200 páginas); Expert \$999/mes (500 páginas); Team \$49/usuario/mes (5,000 páginas); Enterprise: precio personalizado	Suscripción mensual por usuario	Límite de páginas según plan; adicional desde \$0.01–\$0.02/página
SciSpace Copilot	Premium \$12/mes anual (\$20/mes pago mensual); Labs/Univ. \$8/usuario/mes; Free limitado	Suscripción mensual/anual	Premium: export y uso ilimitado; Free: limitado en mensajes y búsquedas
AWS Bedrock (Knowledge Bases)	No precio fijo; se paga por uso (modelos + almacenamiento): p. ej. embeddings ~\$9/mes; vector DB ~\$691/mes; total puede rondar \$766/mes en ejemplo	Pago por uso (tokens, KB, servicios AWS)	No hay límite de páginas; depende de capacidad de KB y costo asociable

Notas relevantes:

- **Microsoft Copilot:** Puede variar según licenciamiento de M365.
- La mayoría de las soluciones **no cobran por página**, salvo **Humata**, que sí incluye un límite mensual de páginas con costo adicional por página extra.

- Varias soluciones (Google Gemini, AWS Bedrock, ChatGPT Enterprise) siguen un modelo “**pay-as-you-go**” según cantidad de tokens o uso, en vez de fijar límite de páginas.
- Nube (SaaS): precios claros (suscripción por usuario), con límites de archivo o uso por plan. Muchos ofrecen niveles academic/student o add-on económicos (como Adobe).

5. Análisis de Soluciones On-Premise

5.1 Descripción de herramientas revisadas

Solución	¿Qué hace?	Tecnología detrás	Mín. de infraestructura
Ollama	Ejecuta LLMs locales y puedes resumir PDFs (vía script/plugin o RAG)	Motor local para modelos abiertos (Llama 3.x, Mistral, Qwen, DeepSeek); CLI/API	Funciona en Windows 10+/macOS 12+/Linux ; CPU-only posible; recomendable ≥8-16 GB RAM y GPU dedicada p/ modelos medianos. Docker y soporte GPU en Linux.
privateGPT	RAG local sobre PDFs (ingestas, indexa y resume sin enviar datos fuera)	LLM abierto + embeddings + FAISS; todo en local	Corre en CPU; mejora con GPU. Guías para WSL/GPU; rendimiento depende del tamaño de modelo.
llama.cpp (motor)	Inferencia local de modelos GGUF para resumir/QA	Backend C++ (GGML/GGUF) para Llama/Mistral/Qwen, etc.	CPU-only viable; para 7B se sugiere ~6-8 GB RAM/VRAM ; 13B ~10+ GB ; mejores resultados con GPU moderna. (rangos de la comunidad).
IBM watsonx (on-prem en OpenShift)	Despliegue on-prem de modelos/LLMs para tareas como resumen	watsonx.ai + Red Hat OpenShift en tu datacenter	Requiere clúster OpenShift; sizing según modelo/uso; opción totalmente on-prem.
NVIDIA NeMo / NIM	Pila on-prem para LLMs y RAG (incl. sobre PDFs)	Inference Microservices (NIM), NeMo, <i>retrievers</i> ; integra GPUs NVIDIA	Servidores con GPU NVIDIA (VRAM según modelo); despliegue local/air-gapped posible.
Haystack (deepset)	Framework open-source para RAG/QA/resumen de documentos	Python + conectores + FAISS/ES/Weaviate; LLM abierto/local	Corre en servidores x86 estándar; CPU-only posible; GPU acelera.

Requisitos mínimos (para on-prem/local)

- **CPU-only (piloto / infraestructura vieja):** se puede correr un modelo **7B** cuantizado (por ejemplo con **llama.cpp/Ollama**) en un servidor con **≥8-16 GB** de RAM; será más lento pero sirve para POC.
- **GPU recomendada (mejor experiencia):** para modelos **~7B-10B** basta **~8-12 GB VRAM**; **30B** suele pedir **~16 GB VRAM**; **70B** en FP16 puede requerir **>140 GB VRAM** o *sharding/quantización*.
- **Compatibilidad Ollama:** Windows 10+, macOS 12+, Linux, con opción Docker y GPU en Linux

5.2 Cuadro Costos, Modalidad de Cobro y Límites

Solución	Modalidad de Cobro / Costo	Límite de Procesamiento / Páginas
Ollama	Open-source, gratuito, sin suscripción	Depende del hardware y tamaño de contexto; sin límite explícito
privateGPT	Open-source, gratuito	Depende de la capacidad local (RAM, CPU/GPU)
llama.cpp	Open-source, gratuito (librería/motor)	Sin límite; depende del hardware y tamaño del contexto
IBM watsonx	Suscripción enterprise. Ejemplo: Standard USD 1,050/mes por 2,500 CUH	Medido por tokens o CUH; no por páginas
NVIDIA NeMo / NIM	Parte de NVIDIA AI Enterprise. Ejemplo: NIM ≈ USD 4,500/anual por GPU	Depende de la capacidad del GPU; no límite de páginas
Haystack (deepset)	Open-source gratuito. Enterprise con precios personalizados. Free Studio: 50 archivos (10 MB c/u) y 100 pipeline hours	Limitado en la versión gratuita; Enterprise sin límite explícito

Notas relevantes:

- **Herramientas totalmente open-source (Ollama, privateGPT, llama.cpp):** Sin costo económico, pero el límite real lo impone el hardware del usuario. Ideal para pilotos rápidos o industrias con alta privacidad donde se prefiere control total.
- **Enterprise on-prem (IBM watsonx, NVIDIA, Haystack Enterprise):** Sí tienen precios definidos o personalizados según uso, infraestructura o soporte requerido. El límite de “páginas” no se maneja de forma explícita, sino que se mide en tokens, horas de capacidad o GPU.
- Estos modelos son flexibles, potentes al maximizar control y privacidad, pero requieren inversión técnica para dimensionar hardware y entornos on-prem.

6. Comparativo (Nube vs. On-Premise)

6.1. Cuadro comparativo General

Criterio	Soluciones en la Nube	Soluciones On-Premise / Infra Cliente
Ejemplos	Adobe Acrobat AI, Microsoft Copilot, Google Vertex, AWS Bedrock	Ollama, Llama2/Mistral locales, LangChain + modelos open source
Facilidad de uso	Muy alta, se consumen vía suscripción o API	Media, requieren instalación y configuración
Parsing de documentos	Integrado en la mayoría (Adobe, Copilot)	Depende de librerías adicionales (Tika, PDFMiner, OCR)
Escalabilidad	Escala masiva en minutos	Limitada por hardware del cliente
Personalización	Limitada (enfocada a casos estándar)	Alta (puede entrenarse/adaptarse a dominio específico)
Seguridad y privacidad	Dependencia de la nube → riesgo de compliance y datos sensibles	Control total del cliente sobre sus datos
Infraestructura mínima	Solo acceso a internet	Servidores con GPU/CPU robustos, RAM alta, almacenamiento
Costos	Pago por suscripción / consumo	Inversión inicial en hardware + soporte
Diferencial clave	Rapidez y simplicidad	Control, privacidad y adaptación

¿Qué tecnología hay detrás (patrones comunes)?

- **LLMs** (modelos base): abiertos (Llama 3.x, Mistral, Qwen, DeepSeek) u ofrecidos como servicio (GPT-4.x/4.1, Claude, Gemini). En cloud se orquesta vía **OpenAI/Azure OpenAI, Anthropic, Vertex AI (Gemini), Amazon Bedrock**, etc.
- **RAG** (Retrieval-Augmented Generation): indexa PDFs y “inyecta” pasajes relevantes al prompt para resúmenes y Q&A con mayor precisión (p. ej., **Bedrock Knowledge Bases** y **Vertex AI Search**).
- **Embeddings + vector DB**: FAISS/ES/Weaviate o servicios gestionados; esencial para buscar fragmentos antes del resumen. (Ver *Knowledge Bases* y documentación de RAG).
- **Parsing de PDFs**: los servicios cloud ya incluyen parseo de tablas/figuras (p. ej., Bedrock KB) y Gemini “document understanding”. En on-prem se suele usar PyMuPDF, unstructured, pdfminer, etc.

6.2. Cuadro comparativo costos

Tipo de Solución	Costos	Forma de Pago	Límites
Nube (Adobe, Copilot, Google, etc.)	Bajo a medio (USD 10–30/mes por usuario; pago por uso en AWS/Vertex)	Suscripción mensual/anual o por tokens	Límite de tokens, páginas o consultas según plan
On-Premise (Ollama, privateGPT, llama.cpp, Haystack, etc.)	Gratis (open source) o alto (IBM/NVIDIA Enterprise)	N/A o licenciamiento anual	No hay límite de páginas; depende del hardware disponible

6.3. Cuadro comparativo Exportación de resúmenes

Solución	¿Entrega resumen en PDF/Word directamente?	Detalle
Adobe Acrobat AI Assistant	✓ Sí	Permite guardar dentro del PDF el resumen generado.
Microsoft Copilot (Word/365/Edge)	✓ Sí	Inserta el resumen en Word, luego exportable a PDF.
Google Gemini / Vertex AI (Docs/Drive)	✓ Sí	Resumen en Google Docs → exportar a Word o PDF.
Box AI	✗ No nativo	Muestra en interfaz; se copia/pega al documento.
Dropbox AI	✗ No nativo	Muestra en interfaz; se copia/pega.
ChatGPT (OpenAI)	✗ No nativo	Texto en interfaz, debes copiar/pegar o usar plugins de

Solución	¿Entrega resumen en PDF/Word directamente?	Detalle
		terceros.
Perplexity Pro	✗ No nativo	Solo muestra en pantalla, no exporta.
Humata	✗ No nativo	Exportación no incluida, resumen visible en interfaz.
SciSpace Copilot	✗ No nativo	Se consulta en web; debes copiarlo manualmente.
AWS Bedrock (Knowledge Bases)	✗ Depende de integración	Devuelve JSON/texto, necesitas integrarlo a PDF/Word.
Ollama	✗ No nativo	Texto en consola/API; se guarda manualmente en Word/PDF.
privateGPT	✗ No nativo	Igual que Ollama, exportación manual o integración adicional.
llama.cpp	✗ No nativo	Motor puro, solo texto; exportación debe programarse.
IBM watsonx (on-prem)	⚠ Parcial	Permite integraciones para reportes PDF/Word, no directo al usuario final.
NVIDIA NeMo / NIM	⚠ Parcial	Se pueden automatizar pipelines para generar documentos, pero requiere

Solución	¿Entrega resumen en PDF/Word directamente?	Detalle
		configuración.
Haystack (deepset)	✗ No nativo	Texto generado en consola/API; exportación programable.

- La nube domina con soluciones accesibles, fáciles de usar y con escalabilidad inmediata (Adobe, Microsoft, Google, OpenAI).
- Los grandes proveedores ofrecen seguridad y compliance robusto, lo que da confianza a corporativos y entornos regulados.
- Modelos flexibles de pago por uso o suscripción, que se adaptan tanto a usuarios individuales como a empresas.
- En el espacio open source / on-premise (Ollama, llama.cpp, Haystack), se destaca la libertad de personalización y la posibilidad de operar en infraestructuras propias sin depender de terceros.

7. Hallazgos clave

7.1. Fortalezas del mercado actual

- La nube domina con soluciones accesibles, fáciles de usar y con escalabilidad inmediata (Adobe, Microsoft, Google, OpenAI).
- Los grandes proveedores ofrecen seguridad y compliance robusto, lo que da confianza a corporativos y entornos regulados.
- Modelos flexibles de pago por uso o suscripción, que se adaptan tanto a usuarios individuales como a empresas.
- En el espacio open source / on-premise (Ollama, llama.cpp, Haystack), se destaca la libertad de personalización y la posibilidad de operar en infraestructuras propias sin depender de terceros.

7.2. Vacíos y brechas identificadas

- No existe un balance real entre simplicidad cloud y control on-premise; las empresas deben elegir entre conveniencia o soberanía de datos.
- Limitada personalización en español: la mayoría de soluciones están orientadas a inglés u otros idiomas dominantes.
- La exportación de resúmenes aún es poco fluida; en la mayoría de casos requiere copiar/pegar o integraciones adicionales.
- Las soluciones on-premise carecen de soporte “listo para usar”; requieren inversión en servidores y personal capacitado.
- En los modelos cloud, los límites de procesamiento de documentos (tokens, páginas o volumen de consultas) generan fricciones en casos de alto uso.
- Costos elevados en enterprise (AWS, Box, Microsoft Copilot) que excluyen a organizaciones medianas o con presupuestos ajustados.

8. Oportunidad para dIAAna Insights

8.1. Posicionamiento potencial

DIANA Insights puede ubicarse como una solución puente entre la nube y el on-premise, ofreciendo la flexibilidad que hoy no entregan los gigantes ni las alternativas open source por separado. Nuestra propuesta se puede centrar en:

- Ser la opción confiable para empresas medianas y grandes que buscan control de datos, pero sin perder la facilidad de uso de la nube o en sus premisas.
- Posicionarse como un aliado estratégico para organizaciones en proceso de transformación digital, ofreciendo un producto adaptable a distintos niveles de infraestructura.
- Integrar parsing avanzado para mejorar la calidad del resumen y reducir pasos manuales.

8.2. Diferenciales estratégicos

A. Adaptabilidad / Personalización

Mientras que Adobe o Dropbox entregan “lo que hay”, DIANA puede ajustarse a:

- Casos de uso específicos (contratos, informes financieros, políticas públicas).
- Integraciones personalizadas con sistemas propios del cliente.
- Privacidad y control de datos

B. DIANA Insights puede:

- Correr on-prem.
- No usar datos de clientes para entrenar.
- Cumplir con estándares de ciberseguridad y normativas

C. Soporte humano cercano

- Los grandes players ofrecen solo soporte en inglés y vía foros.
- DIANA puede diferenciarse con soporte experto en español, disponible y cercano, con ingenieros capacitados de DB System.
- Flexibilidad híbrida (cloud u on-premise): un modelo de despliegue que combina conveniencia con soberanía de datos.
- Parsing inteligente integrado: procesamiento optimizado de PDFs y documentos complejos para generar resúmenes más precisos.
- Enfoque en español: personalización contextual y semántica que mejora la relevancia para usuarios y organizaciones de la región.
- Exportación directa: capacidad de entregar los resúmenes en formatos listos (PDF, Word) sin pasos manuales adicionales.

- Esquema de precios flexible: modelo de suscripción complementado con créditos por uso, atractivo para pymes y corporativos.
- Adaptabilidad tecnológica: integración con APIs y modelos de IA existentes, maximizando compatibilidad y escalabilidad.

9. Conclusiones y Recomendaciones

9.1. Síntesis del estudio

- El mercado actual ofrece un abanico de soluciones **cloud** con fuerte posicionamiento (Adobe, Microsoft, Google, OpenAI) y opciones **open source/on-premise** que privilegian el control, aunque requieren infraestructura robusta.
- No existe hoy un **balance real entre simplicidad cloud y control on-premise**: las empresas deben sacrificar usabilidad o soberanía de datos.
- Los precios en soluciones cloud son variables, con límites de uso y modelos de pago que se ajustan mejor a usuarios individuales o grandes corporativos, dejando un vacío en el segmento medio.
- En el ecosistema hispanohablante, la **falta de personalización lingüística** y contextual representa una oportunidad clara para DIANA Insights.

9.2. Casos de uso concretos

No todas las empresas necesitan lo mismo:

Caso de uso	Herramientas que lo cubren	¿DIANA Insights puede competir aquí?
Resumen rápido de documentos PDF y wordt	Adobe, Copilot, Box AI	Sí, es posible. En modalidad on-premise actualmente funciona con documentos en texto (no imágenes). En modalidad cloud, sí es capaz de procesar ambos tipos)
Análisis de reportes financieros	ChatGPT, Bedrock	No (on premise)
Infraestructura vieja sin nube	Ollama, privateGPT	¿Soportamos on-prem? Si

9.3. Lineamientos sugeridos para avanzar

A. Definir claramente si DIANA puede:

- **Leer PDFs largos y complejos** (no solo texto plano sino tablas, imágenes, gráficos).
 - Con modelos desplegados on-premise, actualmente procesamos texto plano y tablas. En modalidad cloud bajo suscripción, es posible

incluir también análisis de imágenes y gráficos, dependiendo del modelo contratado.

- **Resumir** (extractivo vs. abstractive).
 - Sí, depende del modelo que use el core. En despliegue on-premise, las capacidades se limitan a abstracciones básicas y a las extracciones solicitadas. En modalidad cloud de pago, las capacidades se amplían según el modelo contratado.
- **Hacer Q&A interactivo** sobre el PDF (ejemplo: “¿qué conclusiones tiene la sección 3?”).
 - Sí, depende del modelo que utilice el core. En despliegue on-premise, las capacidades pueden estar limitadas (aunque es posible mejorarlas mediante RAG). En modalidad cloud de pago, estas capacidades se amplían según el modelo contratado.
- **Soportar varios idiomas** (ej: español/inglés) en un mismo documento.
 - Sí.

B. Seguridad y compliance

En los cuadros comparativos, las soluciones Cloud siempre resaltan **residencia de datos, cifrado, privacidad**. Para que DIANA destaque, sería clave explicar:

- Si los datos de los PDFs **no se usan para entrenar** futuras soluciones.
 - Esto es garantizado siempre y cuando se despliegue on premise.
- Qué **controles de seguridad** maneja (cifrado, anonimización, logs).
 - Logs.
- Si soporta **ambientes cerrados (air-gapped / on-premise)**.
 - Ambos.

C. Requisitos de infraestructura (para clientes on-prem)

Tener claridad en lo mínimo:

- **CPU y RAM requeridos**
- Si aprovecha **GPU opcional** para acelerar.
- Compatibilidad con sistemas operativos (Windows/Linux). Esto es importante porque varios clientes tienen infraestructura vieja.

D. Futuras preguntas que nos pueden hacer sobre DIANA

1. ¿Qué motor LLM usa DIANA (propio, open-source, OpenAI, Azure, etc.)?
 - El modelo que se desee integrar, ya sea un open source, propio o de pago cloud.
2. ¿Cómo maneja parsing de PDFs?
 - Solo se puede hacer con documentos digitalizados (no imágenes en PDFs) mediante un sistema RAG.
3. ¿Tiene RAG y vector DB detrás?
 - Si es necesario.
4. ¿Puede correr en on-prem o solo en nube?
 - En ambos, cuando se despliega cloud se necesita una máquina virtual.
5. ¿Cuál es el requisito mínimo de hardware?
 - Definir capacidades: La experiencia muestra que modelos on-premise con CPU limitan abstracciones complejas, mientras que modelos en cloud ofrecen capacidades avanzadas multi-modalidad (texto, tablas, imágenes).
 - Seguridad: El despliegue local de los modelos asegura aislamiento total; cloud bajo el modelo de responsabilidad compartida.
 - Infraestructura: La selección del hardware debe alinearse al modelo, volumen de documentos y velocidad esperada. Modelos grandes (>10B parámetros) se benefician mucho de GPUs modernas.
 - Requisitos obligan a planear: No se recomienda usar modelos 20B+ solo en CPU por lentitud, además debe considerarse memoria RAM $\geq 64\text{GB}$ para esos casos.

Cuadro comparativo: Requisitos mínimos de hardware para modelos usados en DIANA Insights (on premise con Ollama vs. modelos pago Cloud)

Aspecto	Llama 3.1 8B (CPU)	GPT-OSS 20B (CPU/GPU opcional)	Mistral 7B (GPU recomendada)	Modelo pago en cloud (GPT-4, Gemini, Claude)
Parámetros	8 mil millones	20 mil millones	7 mil millones	Variable (16B - 400B +)

Aspecto	Llama 3.1 8B (CPU)	GPT-OSS 20B (CPU/GPU opcional)	Mistral 7B (GPU recomendada)	Modelo pago en cloud (GPT-4, Gemini, Claude)
CPU mínimo	16 cores Xeon/Ryzen recomendados	24+ cores recomendados	16 cores para inferencia ligera	No aplica, servidor cloud manejado por proveedor
RAM mínimo	32-64 GB	64+ GB	32 GB para inferencia	No aplica, recursos escalables en cloud
GPU (opcional/recomendado)	No soporta aceleración GPU	Compatible con GPUs NVIDIA A4000/RTX 3080	Recomendado GPU NVIDIA RTX A4000 o mejor	Cloud usa hardware de última generación
Velocidad (inferencia)	Lenta en CPU, puede ser minutos	Mejor con GPU, lento en CPU	Rápido en GPU	Inmediato, baja latencia
Compatibilidad SO	Linux/Windows (Linux recomendado)	Linux/Windows (Linux preferido)	Linux preferido	N/A
Tipo de modelos soportados	Solo texto, extractivo básico	Texto, abstracción básica	Texto con abstracción avanzada	Multi-modal, capacidades ampliadas
escalabilidad	Limitada a hardware local	Limitada a hardware local	Escalable con más GPUs	Altamente escalable y flexible
Costos	Costos únicos por hardware y licencia	Hardware alto, mantenimiento local	Costos de GPU + licencia/model	Pago por uso (tokens/duración)

9.4. Cuestionario interno – Tecnología detrás de DIANA Insights (DB System) para FICHA TECNICA

1. Modelo y motor de IA

- ¿Qué modelos de lenguaje utiliza DIANA Insights para generar resúmenes?
 - ☒ Modelos propietarios desarrollados por DB System
 - ☒ Modelos abiertos (ej: Llama, Mistral, Falcon, Qwen, etc.)
 - ☒ Modelos comerciales/licenciados (ej: GPT-4, Claude, Gemini, etc.)
- ¿La técnica principal es...?
 - ☒ Generative AI (LLMs)
 - ☒ RAG (Retrieval-Augmented Generation)
 - ☒ Algoritmos extractivos (NLP tradicional)

2. Arquitectura e infraestructura

- ¿DIANA Insights corre en...?
 - ☒ Nube pública (Azure, AWS, GCP u otra)
 - ☒ Nube privada de DB System
 - ☒ On-premise en la infraestructura del cliente
- Si corre on-prem:
 - ¿Qué **sistemas operativos** soporta? (Windows/Linux)

Se recomienda Linux por licencia customización, pero también permite desplegar sobre Windows.

3. Procesamiento de documentos

- ¿Qué tipos de documentos soporta además de PDF?
 - ☒ Word, Excel, PowerPoint
 - ☐ Imágenes escaneadas (OCR integrado)
 - ☒ Otros (especificar)
- ¿Puede interpretar **tablas, gráficos e imágenes** o solo texto?
 - Con modelos On-premise se espera solo interpretación de texto, con modelo Cloud se pueden tener interpretar **tablas, gráficos e imágenes** y solo texto.
- ¿Funciona con documentos en **múltiples idiomas**?
 - Si

4. Seguridad y privacidad

- ¿Qué medidas de seguridad implementa DIANA Insights?
 - ☐ Cifrado en tránsito y en reposo
 - ☒ Logs/auditoría de accesos

- ☒ Control de roles y permisos
 - ☒ Eliminación segura de datos
- ¿Los documentos procesados se usan para **entrenar modelos** o se mantienen aislados del cliente?
 - On-premise se garantiza que mantienen aislados en la estructura del cliente sin ser utilizados con otros fines.

5. Integración y diferenciales

- ¿Con qué sistemas puede integrarse DIANA Insights?
 - ☐ CRMs (ej: Salesforce, Dynamics, etc.)
 - ☐ ERPs (ej: SAP, Oracle, etc.)
 - ☒ Repositorios documentales (SharePoint, Box, Google Drive, etc.)

(Si se tiene acceso a internet en el lugar donde se realiza el despliegue)

- ¿Qué diferenciales aporta frente a un simple “resumen”?
 - ☒ Extracción de KPIs
 - ☒ Dashboards e insights visuales
 - ☐ Modelos predictivos
 - ☒ Recomendaciones accionables