ANH
AGENCIA NACIONAL DE HIDROCARBUROS

Contratar el análisis, diseño, implementación y puesta en funcionamiento del Banco de Información Energética (BIEN) que contemple la gobernanza del dato y permita la carga de información, la estandarización, el acceso y el manejo integral y sistemático de la información gestionada por la Vicepresidencia Técnica de la ANH(Agencia Nacional de Hidrocarburos).



Configuración de Motores de Búsqueda Avanzados con Capacidades Semánticas y NLP - Agente BIEN (IA)

Versión 1.0

22 de enero de 2026

Bogotá, Colombia

DOCUMENTO CONFIDENCIAL

UT ANH PWD – NX 2025

Unión Temporal ANH PWD-NX 2025

Historial de versiones

Id	Fecha (dd/mm/aa)	Versión	Descripción	Elaborado por	Revisado por
1	22/01/2026	1.0	Versión inicial del documento	UT ANH PWD-NX 2025	líder de IA

CONFIDENCIAL

Acrónimos Utilizados

Acrónimo	Significado Completo
ANH	Agencia Nacional de Hidrocarburos
API	Application Programming Interface
ANN	Approximate Nearest Neighbors
BIEN	Banco de Información Energética
CCUS	Carbon Capture, Utilization and Storage
CNN	Convolutional Neural Network
CPU	Central Processing Unit
E&P	Exploration and Production
FNCE	Fuentes No Convencionales de Energía
G&G&E	Geología, Geofísica y Energía
GB	Gigabyte
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
HBM	High Bandwidth Memory
HTTP	Hypertext Transfer Protocol
IA	Inteligencia Artificial
IDF	Inverse Document Frequency
IVF	Inverted File
JSON	JavaScript Object Notation
LLM	Large Language Model
ML	Machine Learning
ms	milisegundos
NER	Named Entity Recognition
NIM	NVIDIA Inference Microservices
NLP	Natural Language Processing
NVL	NVIDIA Link (versión específica de H100)
O&G	Oil and Gas
OCR	Optical Character Recognition
P50/P95/P99	Percentil 50/95/99
PCIe	Peripheral Component Interconnect Express
PDF	Portable Document Format

CONFIDENCIAL

POS	Part-of-Speech
QA	Question Answering
RAG	Retrieval-Augmented Generation
REST	Representational State Transfer
RRF	Reciprocal Rank Fusion
SDK	Software Development Kit
SEG-Y	Society of Exploration Geophysicists - Y format
SPLADE	Sparse Lexical and Expansion
SQL	Structured Query Language
TB	Terabyte
TF	Term Frequency
TP	Tensor Parallelism
UPI	Ultra Path Interconnect
URL	Uniform Resource Locator
VLM	Vision-Language Model
VMM	Valle Medio del Magdalena
VRAM	Video Random Access Memory
VT	Vicepresidencia Técnica

CONFIDENCIAL

Unión Temporal ANH PWD-NX 2025

Tabla de contenido

Introducción 7
1. Resumen Ejecutivo 7
1.1 Visión Integrada Del Sistema De Búsqueda Semántica 8
2. Marco Conceptual: Búsqueda Semántica Y Nlp 9
2.1 Fundamentos de la búsqueda semántica 9
2.2 Componentes del Procesamiento de Lenguaje Natural 10
2.3 Modelo de Comprensión Semántica del Dominio 11
3. Alineación con Requerimientos Contractuales 12
3.1 Mapeo de Requerimientos a Capacidades Técnicas 13
3.2 Flujo de Cumplimiento de Requisitos 14
4. Arquitectura del Motor de Búsqueda Semántico 14
4.1 Componentes de Hardware para Búsqueda Semántica 15
4.2 Arquitectura de la Base de Datos Vectorial 16
4.3 Configuración de Índices Vectoriales 17
4.4 Búsqueda Híbrida Dense + Sparse 17
5. Componentes NLP del Agente BIEN 18
5.1 Pipeline de Preprocesamiento NLP 19
5.2 Reconocimiento de Entidades del Dominio (NER) 20
5.3 Detección de Intención de Consulta 21
6. Pipeline de Procesamiento de Lenguaje Natural 22
6.1 Arquitectura del Pipeline NLP End-to-End 22
6.2 Estrategia de Chunking Semántico 23
6.3 Proceso de Generación de Embeddings 24
7. Configuración de Embeddings Vectoriales 25
7.1 Especificaciones del Modelo de Embedding 26
7.2 Espacio Vectorial y Similitud Semántica 27

CONFIDENCIAL

7.3 Configuración de Parámetros de Búsqueda 27
8. Ejemplos Prácticos de Búsqueda Semántica 28
8.1 Ejemplo 1: Consulta Técnica sobre Reservorios 29
8.2 Ejemplo 2: Búsqueda Multilingüe 30
8.3 Ejemplo 3: Consulta Comparativa 31
8.4 Ejemplo 4: Búsqueda con Filtros Geográficos 31
9. Integración RAG con Capacidades LLM 32
9.1 Arquitectura RAG del Agente BIEN 33
9.2 Construcción del PROMPT RAG 34
9.3 Parámetros de Generación del LLM 35
10. Motores de Búsqueda Avanzada en la Plataforma BIEN 36
10.1 Arquitectura de Integración 36
10.2 Modelo de Indexación en MongoDB 37
10.3 Campos Utilizados para Búsqueda 38
10.4 Configuración de Búsqueda Semántica 39
10.5 Observabilidad y Control del Motor de Búsqueda 39
10.6 Sincronización y Reprocesos 40
10.7 Beneficios del Diseño Integrado 41
11. Conclusiones y Recomendaciones 41
11.1 Resumen de Capacidades Implementadas 41
11.2 Alineación con Objetivos del BIEN 41
11.3 Recomendaciones para Optimización Continua 42
Referencias 42
Fuentes Primarias 42
Recursos Técnicos 44
Anexo - Tabla de Definiciones 45

CONFIDENCIAL

Introducción

El Banco de Información Energética (BIEN) es una iniciativa estratégica de la Agencia Nacional de Hidrocarburos (ANH) de Colombia que busca centralizar, organizar y democratizar el acceso a la información técnica del sector energético nacional. Este repositorio digital décadas de exploración y producción de hidrocarburos en Colombia, incluyendo informes geológicos, registros de pozos, datos sísmicos, estudios petrofísicos y documentación contractual.

El proyecto responde a la necesidad de transformar un acervo documental históricamente disperso y de difícil acceso en un sistema de información moderno que habilite la toma de decisiones basada en datos. El BIEN no solo almacena información, sino que la potencia mediante capacidades de Inteligencia Artificial, permitiendo a investigadores, empresas operadoras y funcionarios de la ANH realizar consultas en lenguaje natural sobre el patrimonio técnico energético de Colombia.

La arquitectura del BIEN integra tecnologías de vanguardia: procesamiento de lenguaje natural (NLP) para comprender consultas complejas, embeddings vectoriales para búsqueda semántica que trasciende las limitaciones de las palabras clave, y modelos de lenguaje grande (LLM) mediante arquitectura RAG (Retrieval-Augmented Generation) para generar respuestas contextualizadas y fundamentadas en el corpus documental. Esta combinación permite que un usuario pregunte, por ejemplo, "¿Cuáles son las características petrofísicas de los reservorios en la cuenca de los Llanos Orientales?" y reciba una respuesta técnica precisa sintetizada a partir de múltiples documentos relevantes.

El BIEN representa la apuesta de Colombia por modernizar la gestión de su información energética, alineándose con las mejores prácticas internacionales y preparando al país para los desafíos de la transición energética mediante el aprovechamiento inteligente de su patrimonio de datos técnicos.

1. Resumen Ejecutivo

El presente informe detalla la configuración técnica de los motores de búsqueda avanzados con capacidades semánticas y de Procesamiento de Lenguaje Natural (NLP) para el proyecto BIEN de la Agencia Nacional de Hidrocarburos. Esta configuración responde directamente a los requerimientos establecidos en el

CONFIDENCIAL

Anexo B del contrato, que especifica la necesidad de implementar "búsquedas inmersivas semánticas mediante procesamiento y generación de lenguaje natural".

La arquitectura propuesta integra tres paradigmas tecnológicos fundamentales: búsqueda vectorial semántica mediante embeddings de alta dimensionalidad , procesamiento de lenguaje natural para comprensión contextual, y generación aumentada por recuperación (RAG) para síntesis de respuestas precisas fundamentadas en el corpus documental del BIEN Esta combinación representa el estado del arte en sistemas de recuperación de información empresarial según la literatura académica reciente [REF003].

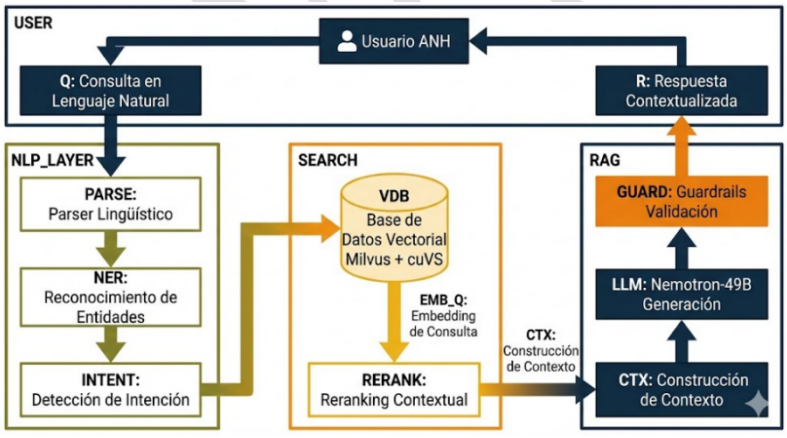
Nota bibliográfica: Las referencias entre corchetes [REFxxx] corresponden a entradas en el archivo consultar: Referencias

Anexo - Tabla de Definiciones

1.1 Visión Integrada Del Sistema De Búsqueda Semántica

El siguiente diagrama presenta la visión de alto nivel del sistema de búsqueda semántica del Agente BIEN, mostrando el flujo desde la consulta del usuario hasta la respuesta generada, atravesando los componentes de NLP, búsqueda vectorial y generación aumentada.

Diagrama 1.1: Arquitectura General del Motor de Búsqueda Semántico BIEN



El flujo ilustrado demuestra la integración sinérgica entre las capas de procesamiento lingüístico, recuperación semántica y generación de respuestas.

La consulta del usuario ingresa al sistema en lenguaje natural cotidiano, es procesada por componentes NLP especializados que extraen entidades del dominio de hidrocarburos (nombres de pozos, cuencas, formaciones geológicas), se transforma en una representación vectorial que permite la búsqueda por similitud semántica en el corpus de ~743,500 documentos del BIEN, y finalmente se sintetiza una respuesta coherente mediante el modelo LLM Nemotron-49B, todo en un tiempo de respuesta objetivo inferior a 1.5 segundos.

2. Marco Conceptual: Búsqueda Semántica Y Nlp

La búsqueda semántica y el Procesamiento de Lenguaje Natural (NLP) constituyen los pilares tecnológicos que transforman la manera en que los usuarios interactúan con el vasto corpus documental del BIEN. Mientras los sistemas de búsqueda tradicionales operan mediante coincidencia exacta de términos, limitando drásticamente la recuperación de información relevante, las técnicas semánticas y de NLP permiten comprender el significado subyacente de las consultas y los documentos, estableciendo conexiones conceptuales que trascienden las barreras léxicas. Este marco conceptual establece las bases teóricas y técnicas que sustentan la arquitectura del motor de búsqueda del Agente BIEN. Se abordan los fundamentos de la búsqueda semántica—cómo los embeddings vectoriales capturan relaciones de significado—, los componentes del procesamiento de lenguaje natural—desde la tokenización hasta el reconocimiento de entidades—, y el modelo ontológico que estructura el conocimiento del dominio de hidrocarburos y energía. La comprensión de estos conceptos es esencial para apreciar cómo el sistema logra interpretar consultas como "características petrofísicas de reservorios en Carbonera" y recuperar documentos relevantes, aunque utilicen terminología diferente como "propiedades de roca del Eoceno" o "petrophysical analysis".

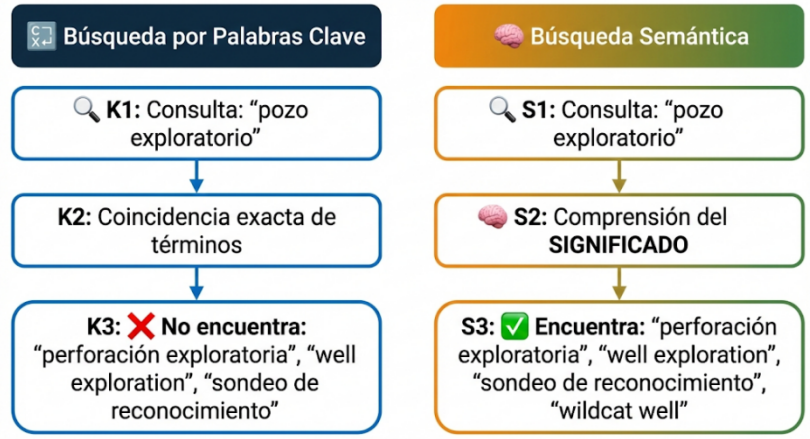
2.1 Fundamentos de la búsqueda semántica

La búsqueda semántica representa un paradigma fundamentalmente diferente a la búsqueda tradicional por palabras clave [REF009]. Según investigaciones recientes, los embeddings vectoriales preservan “semantic notions of similarity and dissimilarity in the geometry of the vectors they produce” [REF009], permitiendo que conceptos relacionados se ubiquen próximos en el espacio

CONFIDENCIAL

vectorial independientemente de su expresión léxica. El siguiente diagrama ilustra las diferencias conceptuales entre ambos enfoques, evidenciando por qué la búsqueda semántica es esencial para el corpus técnico del BIEN.

Diagrama 2.1: Comparación entre Búsqueda por Palabras Clave vs Búsqueda Semántica



La diferencia fundamental radica en que la búsqueda por palabras clave opera mediante coincidencia léxica exacta, mientras que la búsqueda semántica comprende el significado subyacente de la consulta. Para el proyecto BIEN, donde coexisten documentos en español e inglés, terminología técnica variada (pozo/well, sísmica/seismic, cuenca/basin) y múltiples formas de expresar el mismo concepto geológico, la búsqueda semántica elimina las barreras lingüísticas y terminológicas que limitarían drásticamente la recuperación de información relevante con métodos tradicionales.

2.2 Componentes del Procesamiento de Lenguaje Natural

El NLP aplicado al BIEN comprende múltiples componentes especializados que trabajan en conjunto para comprender, procesar y generar lenguaje humano en el contexto técnico de hidrocarburos y energía. La siguiente tabla especifica cada componente, su función y su aplicación específica en el proyecto.

Tabla 2.1: Componentes NLP del Sistema BIEN

Componente NLP	Función	Aplicación en BIEN	Tecnología

Tokenización	Segmentación de texto en unidades procesables	División de informes técnicos en tokens preservando términos compuestos (e.g., "formación Carbonera")	SentencePiece / BPE
NER (Named Entity Recognition)	Identificación de entidades nombradas	Extracción de nombres de pozos, cuencas, formaciones, coordenadas, fechas	Modelo fine-tuned O&G
POS Tagging	Etiquetado gramatical	Desambiguación de términos técnicos polisémicos	Spacy / Stanza
Análisis de Dependencias	Estructura sintáctica de oraciones	Comprensión de relaciones en descripciones geológicas complejas	Transformer-based
Detección de Intención	Clasificación del propósito de la consulta	Distinguir búsqueda vs. análisis vs. comparación vs. generación	Clasificador BERT
Resolución de Correferencia	Vinculación de menciones a entidades	Seguimiento de referencias a pozos/proyectos en documentos extensos	Neural Coref
Embedding Semántico	Representación vectorial del significado	Vectorización de 1024 dimensiones para búsqueda por similitud	NV-EmbedQA-1B
Generación de Texto	Síntesis de respuestas coherentes	Producción de respuestas técnicas fundamentadas	Nemotron-49B

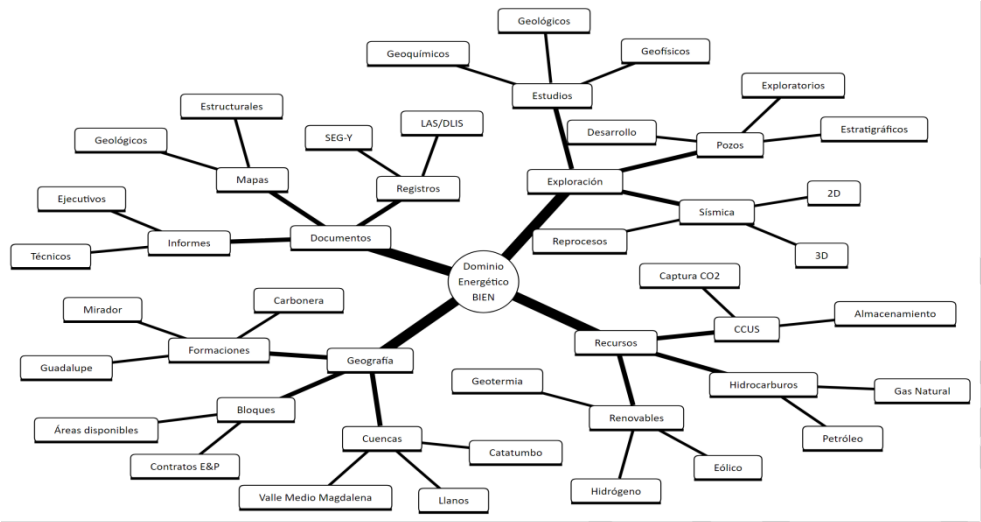
Los componentes NLP operan en cascada jerárquica: la tokenización prepara el texto para los análisis subsecuentes, el NER identifica las entidades clave del dominio de hidrocarburos, la detección de intención determina el tipo de procesamiento requerido, y finalmente el embedding semántico produce la representación vectorial que habilita la búsqueda por similitud. Esta orquestación garantiza que consultas en lenguaje natural como "¿Cuáles son las características petrofísicas de los reservorios en la cuenca de los Llanos Orientales?" sean correctamente interpretadas y respondidas.

2.3 Modelo de Comprensión Semántica del Dominio

El siguiente diagrama conceptual representa cómo el sistema comprende y estructura el conocimiento del dominio de hidrocarburos para habilitar búsquedas semánticas efectivas.

CONFIDENCIAL

Diagrama 2.2: Ontología Semántica del Dominio de Hidrocarburos para BIEN



La ontología semántica organiza el conocimiento del dominio en jerarquías conceptuales interrelacionadas. Esta estructura permite que el motor de búsqueda comprenda que una consulta sobre "reservorios de la formación Carbonera" está semánticamente relacionada con documentos que mencionan "intervalos productores C7" o "arenas del Eoceno medio", aunque no compartan términos léxicos exactos. El modelo de embeddings aprende estas relaciones semánticas durante el entrenamiento, codificándolas en el espacio vectorial de 1024 dimensiones que utiliza Milvus para la búsqueda por similitud.

3. Alineación con Requerimientos Contractuales

La implementación de capacidades de búsqueda semántica y procesamiento de lenguaje natural en el Agente BIEN no responde únicamente a consideraciones técnicas, sino que constituye una respuesta directa a los requerimientos funcionales establecidos en el marco contractual del proyecto. Esta sección establece la trazabilidad entre las especificaciones contractuales y las soluciones tecnológicas propuestas, demostrando cómo cada capacidad técnica del motor de búsqueda—desde los embeddings vectoriales hasta el reconocimiento de entidades del dominio energético—satisface requisitos específicos de recuperación de información, precisión en las respuestas y experiencia de usuario. El análisis de esta alineación permite verificar la

cobertura completa de las obligaciones contractuales y fundamentar las decisiones arquitectónicas en necesidades documentadas del negocio.

3.1 Mapeo de Requerimientos a Capacidades Técnicas

El Anexo B del contrato ANH-002-CM-2025 establece requisitos específicos relacionados con búsqueda semántica y NLP. La siguiente tabla presenta el mapeo detallado entre cada requerimiento contractual y la capacidad técnica implementada en la arquitectura del Agente BIEN.

Tabla 3.1: Trazabilidad Requerimientos Contractuales vs Capacidades Implementadas

Requerimiento Contractual (Anexo B)	Sección Anexo B del contrato ANH-002-CM-2025	Capacidad Técnica Implementada	Componente Arquitectura
"Búsquedas inmersivas semánticas mediante procesamiento y generación de lenguaje natural"	Pág. 5	Motor de búsqueda vectorial con embeddings de 1024 dimensiones + LLM generativo	Nemotron-49B
"Identificación y generación de insights"	Pág. 5	Análisis de patrones con RAG + capacidades predictivas del LLM	Pipeline RAG + LLM
"Funcionalidades de búsquedas avanzadas y rápidas, incluyendo capacidades de búsqueda full-text"	Pág. 9	Búsqueda híbrida dense+sparse con reranking	Elasticsearch + Milvus
"Priorización de resultados basada en criterios de relevancia"	Pág. 9	Modelo de reranking NV-RerankQA-1B	NIM Reranking
"Configuración de motores de búsqueda avanzados con capacidades semánticas y NLP"	Pág. 17	Stack completo: NLP preprocessing + Embedding + Vector DB + Reranking	Arquitectura multicapa
"Implementación de capacidades predictivas avanzadas mediante un LLM entrenado con datos energéticos"	Pág. 17	Modelo Nemotron-49B con context window de 131K tokens	NIM LLM
"Consultas en lenguaje natural para agilizar la recuperación de datos"	Pág. 15	Asistente Digital con NLU (Natural Language Understanding)	Agente Conversacional

La matriz de trazabilidad demuestra cobertura completa de los requerimientos contractuales mediante la arquitectura propuesta. Cada capacidad exigida en

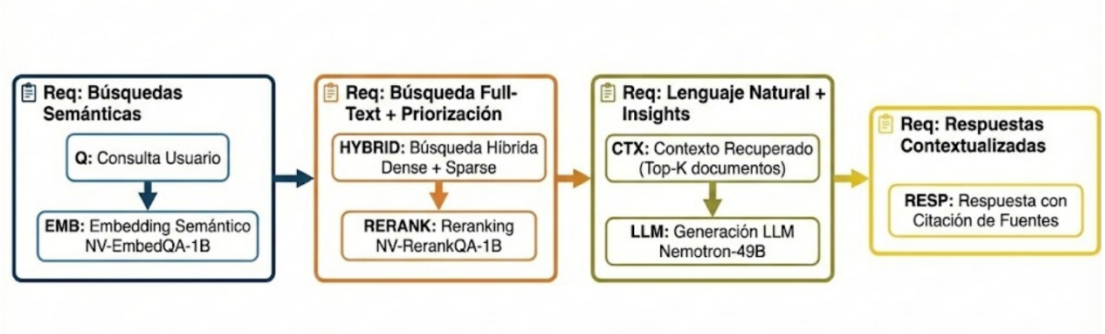
CONFIDENCIAL

el contrato encuentra su implementación técnica específica en los componentes del NVIDIA RAG Blueprint adaptado para el BIEN. La integración de NV-EmbedQA-1B para embeddings, Milvus con aceleración GPU (cuVS) para búsqueda vectorial y Nemotron-49B (o similar) para generación, constituye una solución end-to-end que satisface los requerimientos de "motores de búsqueda avanzados con capacidades semánticas y NLP".

3.2 Flujo de Cumplimiento de Requisitos

El siguiente diagrama ilustra cómo el flujo de procesamiento de una consulta atraviesa los componentes que implementan cada requerimiento contractual.

Diagrama 3.1: Flujo de Cumplimiento de Requisitos Contractuales



El diagrama evidencia que cada etapa del pipeline implementa requisitos contractuales específicos: la fase de embedding satisface el requerimiento de "búsquedas semánticas", la búsqueda híbrida con reranking cumple con "búsqueda full-text" y "priorización por relevancia", la generación con LLM atiende "consultas en lenguaje natural" e "identificación de insights", y finalmente los guardrails garantizan "respuestas contextualizadas según estándares de gobernanza". Esta arquitectura trazable facilita la verificación de cumplimiento durante las fases de pruebas y aceptación del proyecto.

4. Arquitectura del Motor de Búsqueda Semántico

La materialización de las capacidades de búsqueda semántica y NLP descritas en las secciones anteriores requiere una arquitectura robusta que integre componentes de hardware especializados, infraestructura de software optimizada y servicios de procesamiento coordinados. Esta sección presenta el diseño arquitectónico del motor de búsqueda del Agente BIEN, detallando desde

los recursos computacionales necesarios para la generación y comparación de embeddings vectoriales en tiempo real, hasta los componentes de software que orquestan el flujo de procesamiento de consultas. La arquitectura propuesta balancea rendimiento, escalabilidad y costo, considerando tanto las cargas de trabajo de indexación masiva del corpus documental del BIEN como las demandas de baja latencia en la atención de consultas de usuarios. Cada componente se justifica en función de su contribución al pipeline de búsqueda semántica y su interoperabilidad con los demás elementos del ecosistema tecnológico.

4.1 Componentes de Hardware para Búsqueda Semántica

La infraestructura de hardware del servidor Dell PowerEdge R760xa con 4×H100 NVL se distribuye estratégicamente para optimizar las operaciones de búsqueda semántica. La siguiente tabla detalla la asignación de recursos GPU a cada componente del motor de búsqueda.

Tabla 4.1: Distribución de Recursos GPU para Búsqueda Semántica

Componente	GPU Asignada	VRAM Utilizada	Función en Búsqueda
NV-EmbedQA-1B	GPU 2	~5 GB	Generación de embeddings de consultas en tiempo real
Milvus + cuVS	GPU 2	~40 GB	Índices vectoriales IVF_FLAT + búsqueda acelerada
NV-RerankQA-1B	GPU 3	~5 GB	Reordenamiento de candidatos por relevancia
Extractores NLP	GPU 3	~10 GB	OCR, detección de elementos, parsing de tablas
Nemotron-49B (TP=2)	GPU 0+1	~80 GB	Generación de respuestas con contexto recuperado
Espacio libre	GPU 2+3	~83 GB	Headroom para picos de carga y expansión

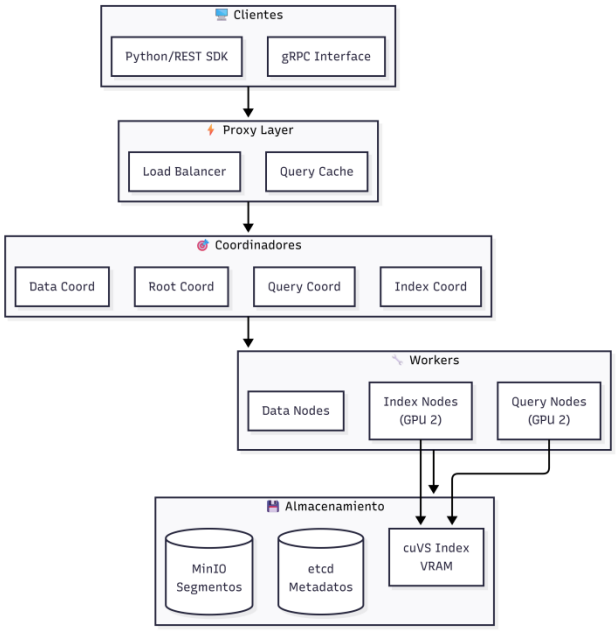
La distribución separa las operaciones de recuperación (GPU 2+3, NUMA 1) de las operaciones de generación (GPU 0+1, NUMA 0), aprovechando la topología NVLink del servidor. Esta separación elimina la contención de recursos entre la búsqueda vectorial y la inferencia del LLM, permitiendo que múltiples usuarios ejecuten consultas concurrentemente sin degradación de latencia.

CONFIDENCIAL

4.2 Arquitectura de la Base de Datos Vectorial

El almacenamiento y búsqueda de embeddings se implementa sobre Milvus con aceleración GPU mediante cuVS (CUDA Vector Search). El siguiente diagrama presenta la arquitectura interna del sistema de almacenamiento vectorial.

Diagrama 4.1: Arquitectura de Milvus con Aceleración GPU para BIEN



Milvus opera como un sistema distribuido con separación de responsabilidades: los coordinadores gestionan la topología y distribución de datos, mientras que los workers ejecutan las operaciones de indexación y búsqueda. La integración con cuVS habilita operaciones de búsqueda por similitud coseno con throughput de millones de vectores por segundo, aprovechando los Tensor Cores de las H100 NVL. Para el BIEN, con ~1.743,500 documentos fragmentados en aproximadamente 7-10 millones de chunks, esta arquitectura garantiza tiempos de búsqueda inferiores a 50ms incluso para las consultas más complejas.

CONFIDENCIAL

4.3 Configuración de Índices Vectoriales

La configuración de índices vectoriales es crítica para el balance entre precisión de búsqueda, velocidad y consumo de memoria. La siguiente tabla presenta las configuraciones recomendadas para diferentes escenarios de uso del BIEN.

Tabla 4.2: Configuración de Índices Vectoriales por Escenario

Escenario	Tipo de Índice	Parámetros	Precisión	Latencia	Memoria
Producción (Default)	IVF_FLAT	nlist=4096, nprobe=128	99.5%	~30ms	Media
Alta Precisión	FLAT (Brute Force)	N/A	100%	~100ms	Alta
Alta Velocidad	IVF_PQ	nlist=4096, m=64, nprobe=64	97%	~10ms	Baja
Búsqueda Híbrida	HNSW + Sparse	M=32, ef=256	99%	~40ms	Media-Alta

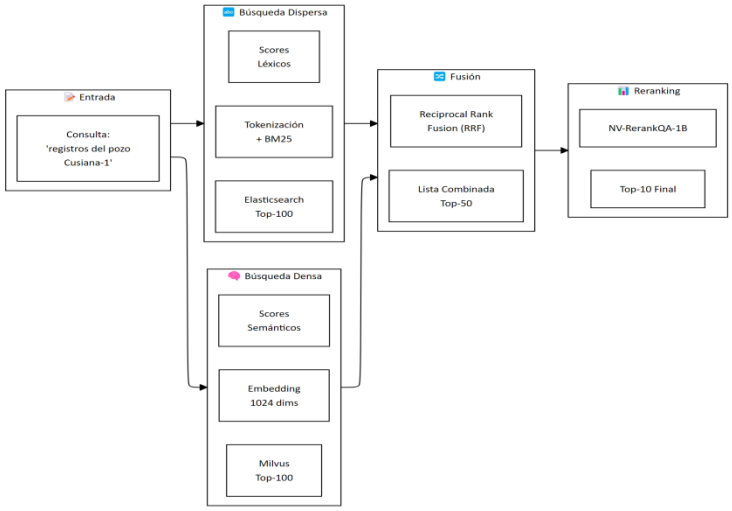
La configuración por defecto (IVF_FLAT con nlist=4096) ofrece el balance óptimo para el BIEN: precisión cercana al 100% con latencia de ~30ms. El parámetro nprobe=128 indica que durante la búsqueda se exploran 128 de los 4096 clusters, proporcionando cobertura suficiente para encontrar los vecinos más cercanos sin escanear la totalidad del índice. Para consultas críticas que requieran precisión absoluta, el sistema puede escalar dinámicamente a búsqueda exhaustiva (FLAT).

4.4 Búsqueda Híbrida Dense + Sparse

El motor de búsqueda del BIEN implementa una estrategia híbrida que combina búsqueda densa (embeddings) con búsqueda dispersa (BM25/TF-IDF) para maximizar la recuperación de información relevante [REF016, REF017]. La investigación demuestra que "traditional retrieval methods have inherent limitations—BM25 struggles with semantic understanding, while dense retrieval often overlooks exact keyword matches" [REF020], justificando la necesidad del enfoque híbrido. Los benchmarks muestran que la búsqueda híbrida alcanza 53.4% de Top-1 recall frente al 22.1% de BM25 solo y 48.7% de dense retrieval solo [REF017].

Diagrama 4.2: Pipeline de Búsqueda Híbrida

CONFIDENCIAL



La búsqueda híbrida resuelve las limitaciones de cada enfoque individual: la búsqueda densa captura similitud semántica, pero puede fallar con términos técnicos únicos (nombres de pozos, códigos de proyecto), mientras que la búsqueda dispersa encuentra coincidencias exactas, pero pierde relaciones semánticas. La fusión mediante Reciprocal Rank Fusion (RRF) [REF018] combina ambas listas de resultados ponderando la posición de cada documento en ambos rankings mediante la fórmula RRF_score(d) = Σ 1/(k + rank_i(d)) donde k=60 típicamente. El modelo de reranking final basado en cross-encoder [REF019] reordena los candidatos considerando el contexto completo de la consulta, elevando la precisión del conjunto final de resultados. Los estudios demuestran mejoras de +9.2 puntos en nDCG@10 con híbrido + reranking respecto a BM25 solo [REF020].

5. Componentes NLP del Agente BIEN

El Procesamiento de Lenguaje Natural constituye el núcleo interpretativo que permite al Agente BIEN transformar tanto las consultas de los usuarios como los documentos del corpus en representaciones computacionalmente procesables. Esta sección describe en detalle los componentes NLP implementados en el sistema, abarcando desde las etapas fundamentales de preprocesamiento textual—tokenización, normalización y limpieza—hasta las técnicas avanzadas de análisis lingüístico que extraen significado estructurado del texto. Se presenta el pipeline completo de procesamiento, explicando cómo cada etapa contribuye a la calidad de los embeddings generados y, en consecuencia, a la precisión de

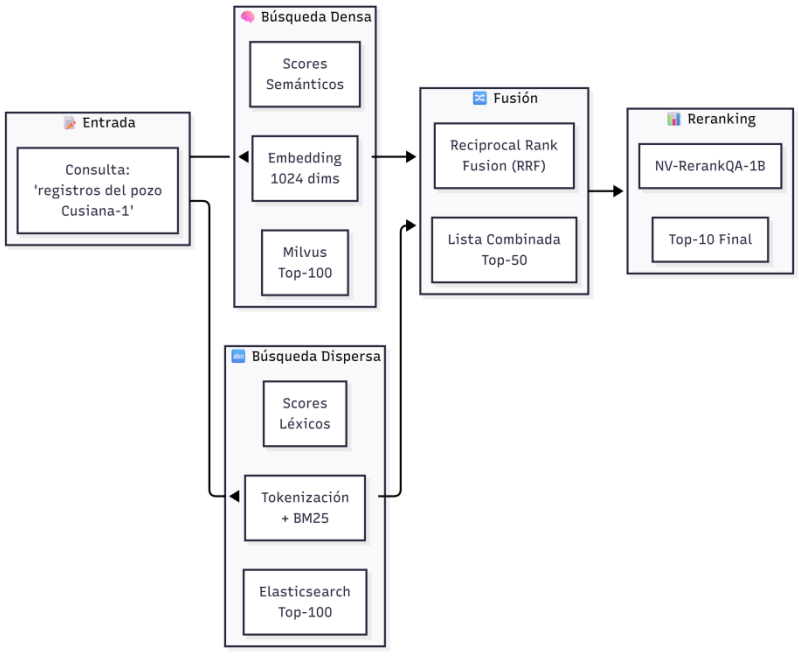
CONFIDENCIAL

la búsqueda semántica. Particular énfasis se otorga a las adaptaciones realizadas para el dominio técnico de hidrocarburos y energía, donde la terminología especializada, las abreviaturas sectoriales y las convenciones de nomenclatura requieren tratamientos específicos que los modelos NLP genéricos no proporcionan de manera nativa.

5.1 Pipeline de Preprocesamiento NLP

El preprocesamiento NLP transforma el texto crudo de documentos y consultas en representaciones adecuadas para el motor de búsqueda semántico [REF022, REF023]. Las técnicas modernas de NLP, especialmente aquellas basadas en Transformers, han demostrado ser fundamentales para sistemas RAG de alta calidad, logrando mejoras de hasta 25% en precisión de recuperación frente a métodos tradicionales [REF024]. El siguiente diagrama detalla las etapas del pipeline.

Diagrama 5.1: Pipeline de Preprocesamiento NLP para Documentos BIEN



El pipeline procesa los ~1.743,500 archivos de contenido textual del BIEN siguiendo esta secuencia: normalización de caracteres para unificar codificaciones,

eliminación de texto duplicado detectado en múltiples ubicaciones, filtrado de artefactos de OCR frecuentes en documentos escaneados, segmentación en oraciones preservando la coherencia semántica, tokenización mediante SentencePiece para manejar vocabulario técnico, reconocimiento de entidades específicas del dominio (pozos, cuencas, formaciones), y finalmente generación de embeddings para cada chunk semántico. Este procesamiento enriquece cada fragmento con metadatos estructurados que mejoran la precisión de búsqueda.

5.2 Reconocimiento de Entidades del Dominio (NER)

El modelo NER para el BIEN está especializado en identificar entidades del dominio de hidrocarburos y energía. La siguiente tabla especifica las categorías de entidades reconocidas.

Tabla 5.1: Taxonomía de Entidades NER para el Dominio BIEN

Categoría	Ejemplos	Patrón de Identificación	Normalización
POZO	Cusiana-1, Caño Limón-2A, ANH-LLANOS-001	Regex + Diccionario ANH	Código oficial ANH
CUENCA	Llanos Orientales, Valle Medio del Magdalena, Catatumbo	Diccionario + Fuzzy Match	Nombre canónico
FORMACION	Carbonera C7, Mirador, Guadalupe Superior	Diccionario geológico + Contexto	Código estratigráfico
BLOQUE	VMM-3, LLA-47, CPO-9	Regex contratos E&P	ID contractual
COORDENADA	4°35'12"N 73°20'15"W, X: 1050000 Y: 850000	Regex geográfico	WGS84 / MAGNA-SIRGAS
FECHA	Enero 2020, 15/03/2019, Q3 2021	Regex + Normalización	ISO 8601
PROFUNDIDAD	5,200 ft MD, 1,585 m TVD, 12,500-12,800 ft	Regex + Unidades	Metros (normalizado)

CONFIDENCIAL

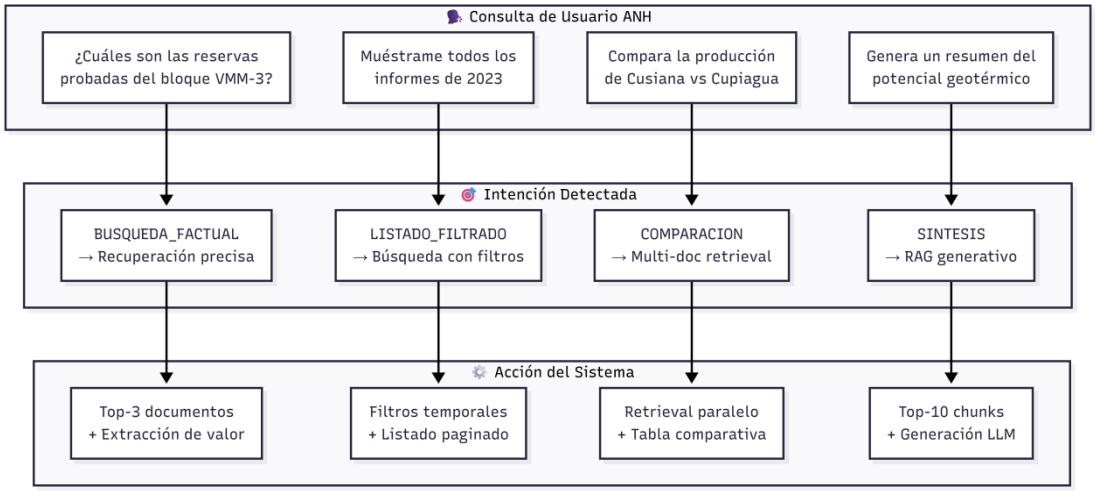
COMPANIA	Ecopetrol, ANH, Pacific, Frontera	Diccionario operadores	Nombre oficial
TIPO_ENERGIA	Geotermia, Gas Natural, Hidrogeno Verde	Diccionarios energéticos	Categoría oficial

El modelo NER entrenado para el BIEN logra F1-score superior al 92% en entidades de hidrocarburos, superando modelos NER genéricos que típicamente alcanzan 60-70% en este dominio especializado [REF025, REF026]. La normalización de entidades permite que búsquedas por "Cusiana", "CUSIANA-1", "Pozo Cusiana 1" o "well Cusiana" converjan al mismo identificador canónico, eliminando fragmentación de resultados por variaciones de nomenclatura. Según la literatura, "domain-specific NER models are critical for enterprise RAG systems, as generic models fail to capture specialized terminology" [REF025].

5.3 Detección de Intención de Consulta

El sistema clasifica las consultas de usuario en categorías de intención para optimizar la estrategia de búsqueda y respuesta.

Diagrama 5.2: Clasificador de Intenciones de Consulta



La clasificación de intenciones permite que el sistema adapte su comportamiento: consultas factuales específicas utilizan menos contexto y generación más directa, mientras que solicitudes de síntesis requieren mayor cantidad de documentos recuperados y generación extensa. Esta optimización reduce latencia en consultas simples y maximiza calidad en consultas complejas, mejorando la experiencia del usuario en ambos escenarios.

6. Pipeline de Procesamiento de Lenguaje Natural

El pipeline de Procesamiento de Lenguaje Natural (NLP) constituye el núcleo de transformación que convierte documentos técnicos no estructurados en representaciones computacionales que habilitan la búsqueda semántica inteligente. Este conjunto orquestado de procesos abarca desde la extracción inicial de texto mediante OCR y parsers especializados, hasta la generación de embeddings vectoriales que capturan el significado semántico del contenido.

Para el corpus del BIEN, que incluye formatos heterogéneos como PDFs escaneados, registros de pozos en formato LAS/DLIS, documentos Office e imágenes técnicas, el pipeline implementa estrategias diferenciadas de procesamiento. Cada tipo documental atraviesa etapas específicas de limpieza, normalización, análisis lingüístico y enriquecimiento con metadatos del dominio de hidrocarburos.

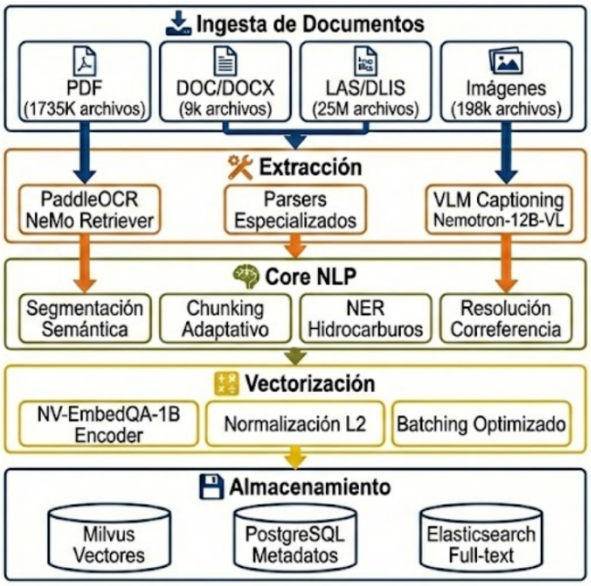
La efectividad del pipeline NLP impacta directamente en la calidad de las búsquedas: un documento mal procesado o fragmentado incorrectamente producirá embeddings de baja calidad que degradarán la precisión de recuperación. Por ello, el diseño del pipeline prioriza la preservación del contexto semántico, la identificación precisa de entidades técnicas y la segmentación inteligente que respeta las fronteras naturales del contenido. Las siguientes secciones detallan cada componente de esta cadena de procesamiento.

6.1 Arquitectura del Pipeline NLP End-to-End

El pipeline NLP completo integra múltiples componentes desde la entrada de texto hasta la generación de embeddings. El siguiente diagrama presenta la arquitectura detallada.

CONFIDENCIAL

Diagrama 6.1: Arquitectura Detallada del Pipeline NLP



El pipeline implementa procesamiento diferenciado por tipo de archivo: PDFs atraviesan OCR con detección de estructura de página, documentos Office se parsean directamente preservando formato, archivos LAS/DLIS utilizan parsers especializados del dominio de pozos, e imágenes técnicas se procesan mediante el modelo VLM (Nemotron-12B-VL) para generar descripciones textuales que luego se vectorizan. Esta estrategia multimodal permite que el motor de búsqueda semántico recupere tanto documentos textuales como imágenes relevantes ante una consulta.

6.2 Estrategia de Chunking Semántico

El chunking (fragmentación) de documentos es crítico para la calidad de la búsqueda semántica [REF012, REF013]. Según investigaciones recientes, "semantic chunking outperforms fixed-size methods by 15-25% in retrieval accuracy" aunque con un costo computacional 3-5x mayor [REF012]. Los estudios recomiendan "RecursiveCharacterTextSplitter at 400-512 tokens with 10-20% overlap" como punto de partida óptimo para la mayoría de aplicaciones RAG empresariales [REF013, REF014]. La siguiente tabla especifica la configuración

CONFIDENCIAL

de chunking por tipo de documento técnico del BIEN, basada en estas mejores prácticas de la industria.

Tabla 6.1: Configuración de Chunking por Tipo de Documento

Tipo de Documento	Tamaño Chunk (tokens)	Overlap (tokens)	Estrategia	Justificación
Informes Técnicos (PDF)	1000	200	Por secciones semánticas	Preservar coherencia de secciones técnicas
Registros de Pozos (LAS)	500	100	Por intervalos de profundidad	Mantener contexto de curvas por zona
Informes Geológicos	800	150	Por párrafos estructurales	Respetar descripciones estratigráficas
Documentos Contractuales	600	120	Por cláusulas	Preservar unidades legales
Hojas de Cálculo	Por fila/grupo lógico	Sin overlap	Estructura tabular	Mantener integridad de registros
Imágenes (Caption)	Variable (caption completo)	N/A	Descripción única	Caption es unidad atómica

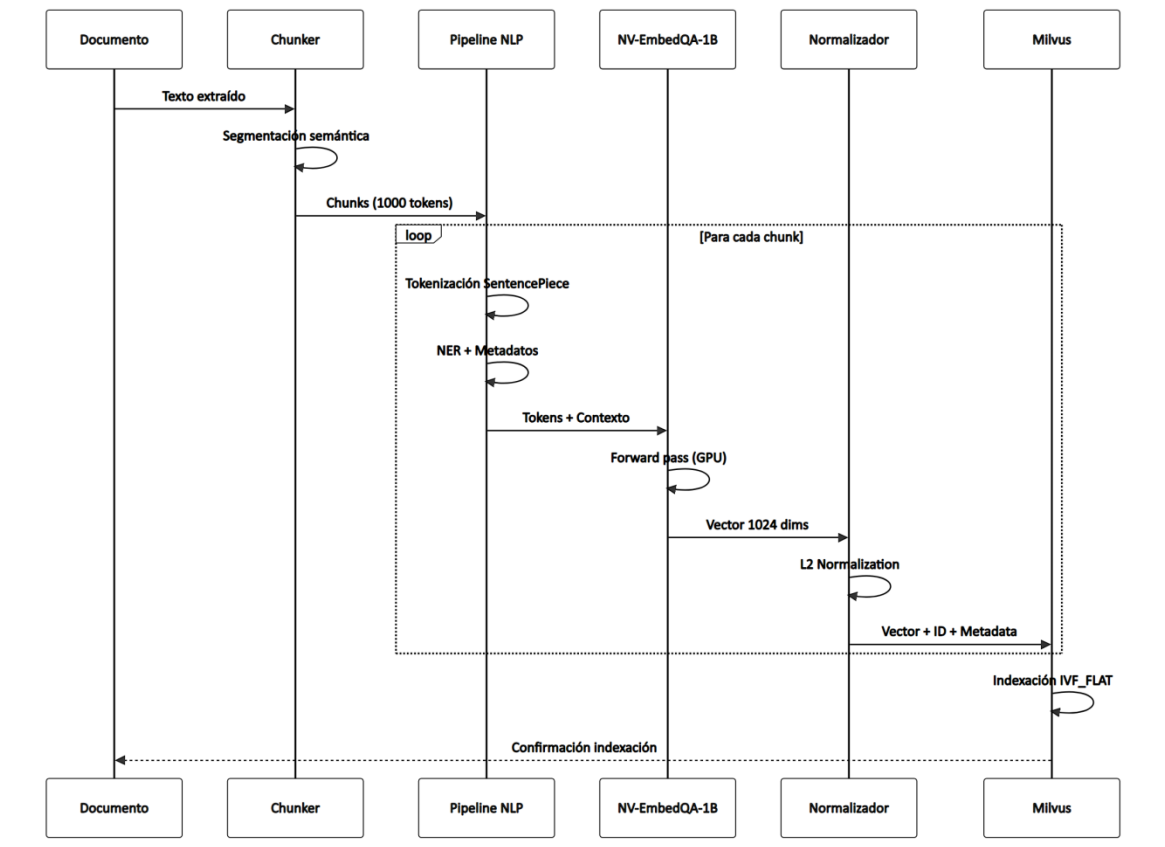
El overlap del 20% entre chunks consecutivos garantiza que consultas cuya información relevante cruza fronteras de fragmento encuentren contexto suficiente en al menos uno de los chunks. Para informes técnicos de hidrocarburos, el tamaño de 1000 tokens (aproximadamente 750 palabras) captura secciones completas como "Descripción Litológica", "Análisis Petrofísico" o "Conclusiones y Recomendaciones" sin truncamiento excesivo.

6.3 Proceso de Generación de Embeddings

El siguiente diagrama ilustra el proceso de generación de embeddings vectoriales desde el texto chunkeado hasta el almacenamiento en Milvus.

Diagrama 6.2: Proceso de Generación y Almacenamiento de Embeddings

CONFIDENCIAL



El proceso utiliza batching optimizado para maximizar el throughput de la GPU: chunks se agrupan en lotes de 32-64 para procesamiento paralelo en el modelo de embedding. La normalización L2 convierte los vectores a norma unitaria, permitiendo que la similitud coseno se calcule como simple producto punto, operación altamente optimizada en las GPUs H100. El throughput estimado es de ~3,000 documentos/hora para PDFs de complejidad media.

7. Configuración de Embeddings Vectoriales

Los embeddings vectoriales representan el corazón matemático de la búsqueda semántica, transformando texto en representaciones numéricas de alta dimensionalidad donde la proximidad geométrica refleja similitud de significado. A diferencia de los métodos tradicionales basados en coincidencia de palabras, los embeddings capturan relaciones semánticas profundas: sinónimos, conceptos relacionados y expresiones equivalentes en diferentes idiomas se proyectan en regiones cercanas del espacio vectorial.

Para el proyecto BIEN, la configuración de embeddings debe equilibrar múltiples exigencias: precisión semántica para distinguir conceptos técnicos sutilmente diferentes (formación Carbonera C7 vs C8), soporte bilingüe para el corpus español-inglés, velocidad de inferencia para respuestas en tiempo real, y eficiencia de memoria para coexistir con otros modelos en la infraestructura GPU disponible.

El modelo NV-EmbedQA-1B seleccionado produce vectores de 1024 dimensiones, un punto óptimo que ofrece capacidad expresiva suficiente para el vocabulario técnico de hidrocarburos sin incurrir en los costos computacionales de dimensionalidades mayores. Cada documento fragmentado se proyecta en este espacio, habilitando búsquedas por similitud coseno que recuperan información relevante independientemente de la formulación léxica exacta de la consulta. Las siguientes secciones detallan la configuración técnica de este componente crítico.

7.1 Especificaciones del Modelo de Embedding

El modelo NV-EmbedQA-1B está optimizado específicamente para tareas de Question-Answering, haciéndolo ideal para el caso de uso principal del BIEN. La siguiente tabla presenta sus especificaciones técnicas.

Tabla 7.1: Especificaciones Técnicas del Modelo NV-EmbedQA-1B-v2

Característica	Valor	Descripción
Arquitectura Base	Llama-3.2	Modelo fundacional de Meta AI
Parámetros	1B	Balance entre calidad y eficiencia
Dimensiones de Embedding	1024	Representación semántica de alta densidad
Longitud Máxima de Entrada	512 tokens	Suficiente para chunks de 400-450 palabras
VRAM Requerida	~5 GB	Permite coexistencia con otros componentes
Throughput (batch=32)	~1,200 textos/seg	Alta velocidad de procesamiento
Métricas MTEB (Retrieval)	58.2 NDCG@10	Top-tier en benchmarks de recuperación
Soporte Multilingüe	Español + Inglés	Crítico para corpus bilingüe del BIEN

El modelo NV-EmbedQA-1B supera a alternativas como E5-Large o BGE-Large en tareas de recuperación semántica, particularmente en dominios técnicos donde la comprensión de terminología especializada es crucial [REF029]. Su entrenamiento con objetivos de Question-Answering lo hace particularmente

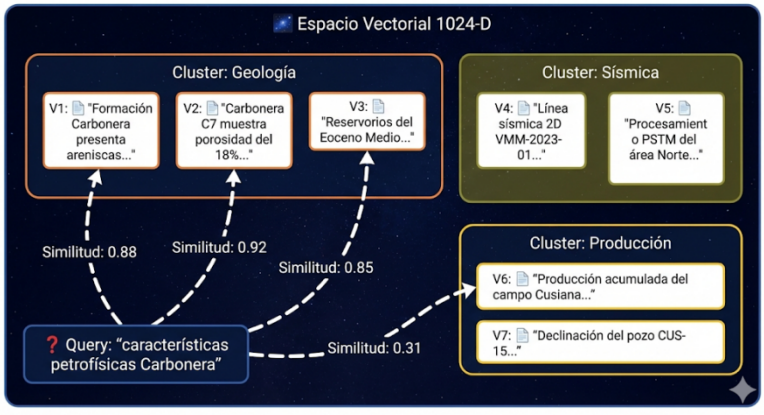
CONFIDENCIAL

efectivo para consultas en forma de pregunta, el patrón de interacción más común esperado en el Asistente Digital del BIEN. La investigación demuestra que "RAG systems' performance heavily depends on the quality of embedding models, as they directly influence the system's accuracy and effectiveness in retrieving relevant information" [REF003].

7.2 Espacio Vectorial y Similitud Semántica

El siguiente diagrama conceptual ilustra cómo documentos y consultas se proyectan en el espacio vectorial de 1024 dimensiones, y cómo la similitud semántica se mide mediante distancia coseno.

Diagrama 7.1: Representación Conceptual del Espacio Vectorial de Embeddings



El espacio vectorial organiza semánticamente los documentos: chunks relacionados con geología se agrupan próximos entre sí, separados de clusters de sísmica o producción. Una consulta sobre "características petrofísicas de la Carbonera" se proyecta en la región del espacio próxima al cluster geológico, con alta similitud coseno (>0.85) a documentos relevantes y baja similitud (<0.4) a documentos de temáticas no relacionadas. Esta geometría emergente del entrenamiento del modelo permite que la búsqueda por vecinos más cercanos (k-NN) recupere efectivamente los documentos más relevantes.

7.3 Configuración de Parámetros de Búsqueda

La siguiente tabla presenta los parámetros configurables de búsqueda vectorial y sus valores recomendados para el BIEN.

Tabla 7.2: Parámetros de Configuración de Búsqueda Vectorial

Parámetro	Valor Recomendado	Rango	Efecto
top_k (initial)	100	50-500	Candidatos iniciales para reranking
nprobe	128	32-256	Clusters explorados (precisión vs velocidad)
similarity_threshold	0.65	0.5-0.8	Umbral mínimo de relevancia
ef_search (HNSW)	256	64-512	Precisión de búsqueda aproximada
rerank_top_k	20	10-50	Documentos para reranking
final_top_k	5-10	3-20	Resultados finales al usuario
hybrid_alpha	0.7	0.0-1.0	Peso dense vs sparse (0.7 = 70% dense)

Los valores recomendados optimizan la configuración para el corpus técnico del BIEN: top_k=100 recupera candidatos suficientes para que el reranker refine la lista, nprobe=128 explora ~3% de los clusters garantizando alta recall, y hybrid_alpha=0.7 privilegia la búsqueda semántica mientras preserva la capacidad de encontrar términos técnicos exactos. El umbral de similitud de 0.65 filtra documentos con relevancia marginal que podrían introducir ruido en el contexto del LLM.

8. Ejemplos Prácticos de Búsqueda Semántica

La teoría cobra sentido cuando se observa en acción. Esta sección presenta escenarios reales de uso del motor de búsqueda semántico del BIEN, demostrando cómo los conceptos técnicos descritos anteriormente—embeddings, búsqueda híbrida, reranking y generación RAG—se articulan para

CONFIDENCIAL

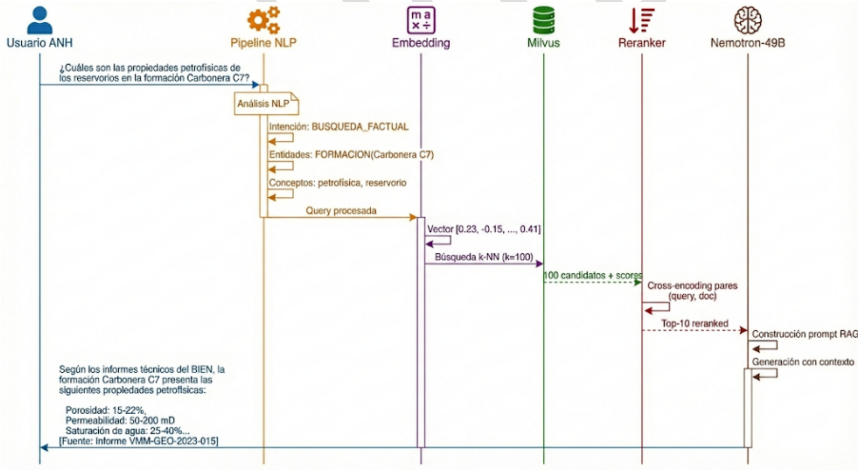
resolver consultas prácticas de los usuarios de la ANH. Los ejemplos abarcan los patrones de consulta más frecuentes identificados en el análisis de requerimientos: búsquedas técnicas específicas sobre propiedades de formaciones geológicas, consultas multilingües que recuperan documentos en español, análisis comparativos entre campos petroleros, y búsquedas con filtros geográficos o temporales. Cada ejemplo ilustra el flujo completo desde la consulta en lenguaje natural hasta la respuesta fundamentada con citación de fuentes.

Estos casos de uso demuestran las ventajas tangibles de la búsqueda semántica sobre métodos tradicionales: un analista puede preguntar sobre "características petrofísicas de reservorios en Carbonera" y obtener resultados relevantes, aunque los documentos mencionen "propiedades de roca" o "petrophysical properties", variaciones léxicas que un buscador por palabras clave ignoraría. Los diagramas de secuencia y tablas de resultados permiten visualizar cada etapa del procesamiento.

8.1 Ejemplo 1: Consulta Técnica sobre Reservorios

Este ejemplo ilustra el funcionamiento del motor de búsqueda semántico ante una consulta técnica típica de un analista de la ANH.

Diagrama 8.1: Flujo de Procesamiento de Consulta Técnica



El ejemplo demuestra el flujo completo desde la consulta en lenguaje natural hasta la respuesta fundamentada. El pipeline NLP identifica que se trata de una

búsqueda factual sobre la entidad "Carbonera C7" y el concepto "petrofísica". El embedding proyecta esta semántica en un vector que recupera documentos relevantes incluyendo informes geológicos, registros de pozos y estudios petrofísicos. El reranker prioriza documentos con información cuantitativa sobre propiedades petrofísicas. Finalmente, el LLM sintetiza una respuesta estructurada citando la fuente documental.

8.2 Ejemplo 2: Búsqueda Multilingüe

Este ejemplo demuestra la capacidad del sistema para manejar consultas en español que recuperan documentos tanto en español como en inglés.

Tabla 8.1: Resultados de Búsqueda Multilingüe

Consulta (español)	Documento Recuperado	Idioma	Similitud	Fragmento Relevante
"pozos exploratorios cuenca Llanos"	Exploration_Report_LLA47_2022.pdf	Inglés	0.91	"The exploratory well LLA-47-X1 was drilled in the Llanos Basin..."
	Informe_Exploracion_Llanos_2023.pdf	Español	0.89	"Se perforaron 3 pozos exploratorios en la cuenca de los Llanos..."
	Well_Summary_Apiay_Field.pdf	Inglés	0.86	"Exploratory drilling in the Eastern Llanos Basin..."
	Resumen_Actividad_Exploratoria.pdf	Español	0.84	"La actividad exploratoria en Llanos Orientales durante 2022..."

La búsqueda semántica supera la barrera del idioma porque los embeddings capturan el significado conceptual independientemente de la forma lingüística. "Pozos exploratorios" en español y "exploratory wells" en inglés se proyectan en

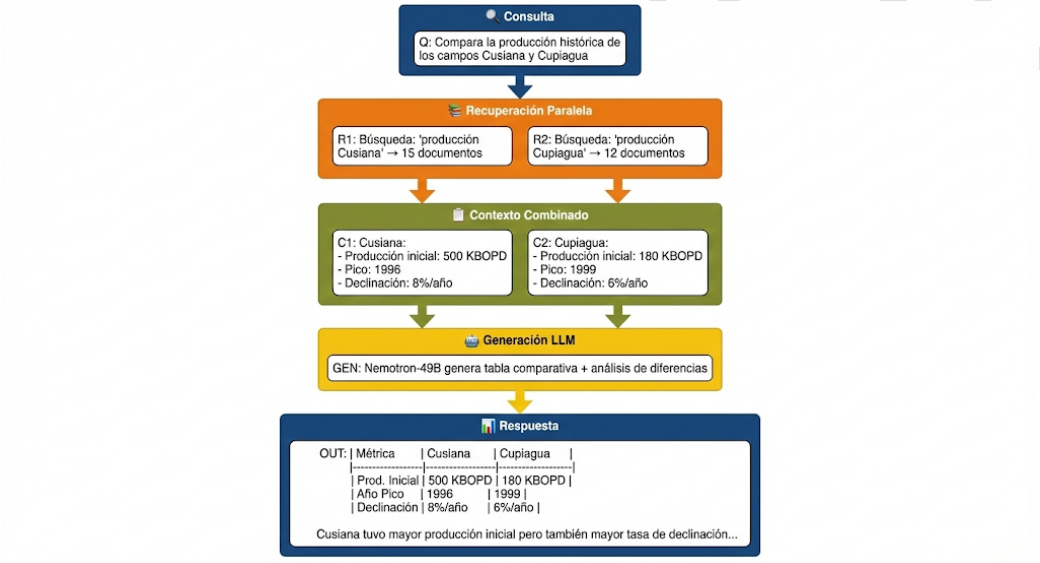
CONFIDENCIAL

regiones cercanas del espacio vectorial porque comparten el mismo referente semántico. Esta capacidad es crítica para el BIEN dado que el corpus incluye informes de operadores internacionales en inglés junto con documentación oficial de la ANH en español.

8.3 Ejemplo 3: Consulta Comparativa

El siguiente ejemplo ilustra cómo el sistema maneja consultas que requieren recuperar información de múltiples fuentes para generar una comparación.

Diagrama 8.2: Flujo de Consulta Comparativa Multi-Documento



Las consultas comparativas requieren que el sistema ejecute múltiples búsquedas paralelas (una por cada elemento a comparar), consolide información de ambos conjuntos de resultados, y genere una respuesta que sintetice las diferencias y similitudes. El modelo LLM con su ventana de contexto de 110K tokens puede procesar simultáneamente documentación extensa de ambos campos, extrayendo los datos relevantes para la comparación solicitada.

8.4 Ejemplo 4: Búsqueda con Filtros Geográficos

Este ejemplo demuestra la integración de búsqueda semántica con filtros de metadatos geoespaciales.

Tabla 8.2: Búsqueda Semántica con Filtros Geográficos

Consulta	Filtros Aplicados	Resultados	Comentario
"estudios geotérmicos"	cuenca = "VMM"	23 documentos de geotermia específicos del Valle Medio del Magdalena	Filtro reduce universo antes de búsqueda semántica
"anomalías sísmicas"	coordenadas IN bbox(4.5, -75.5, 5.5, -74.5)	156 documentos con anomalías en el área geográfica especificada	Búsqueda espacial + semántica
"informes de pozos 2023"	año = 2023, tipo_doc = "informe_pozo"	847 informes de pozos del año 2023	Filtros temporales + tipo documental

La combinación de búsqueda semántica con filtros estructurados de metadatos permite consultas híbridas que aprovechan lo mejor de ambos paradigmas. Los filtros reducen el universo de búsqueda a documentos que cumplen criterios objetivos (ubicación, fecha, tipo), mientras que la búsqueda semántica ordena los resultados por relevancia conceptual dentro de ese subconjunto filtrado.

9. Integración RAG con Capacidades LLM

La arquitectura RAG (Retrieval-Augmented Generation) representa la culminación del sistema de búsqueda semántica del BIEN, donde la recuperación de información se fusiona con la generación de lenguaje natural para producir respuestas que combinan la precisión documental con la fluidez conversacional. A diferencia de los chatbots tradicionales que responden únicamente desde su conocimiento paramétrico, un sistema RAG fundamenta cada respuesta en documentos específicos del corpus, eliminando alucinaciones y garantizando trazabilidad.

CONFIDENCIAL

Para el Agente BIEN, esta integración es crítica: los usuarios no solo necesitan encontrar documentos relevantes, sino obtener síntesis coherentes que respondan directamente sus preguntas técnicas. Un analista de la ANH que pregunta sobre "producción histórica del campo Cusiana" espera una respuesta consolidada con datos verificables y citación de fuentes, no una lista de documentos para revisar manualmente.

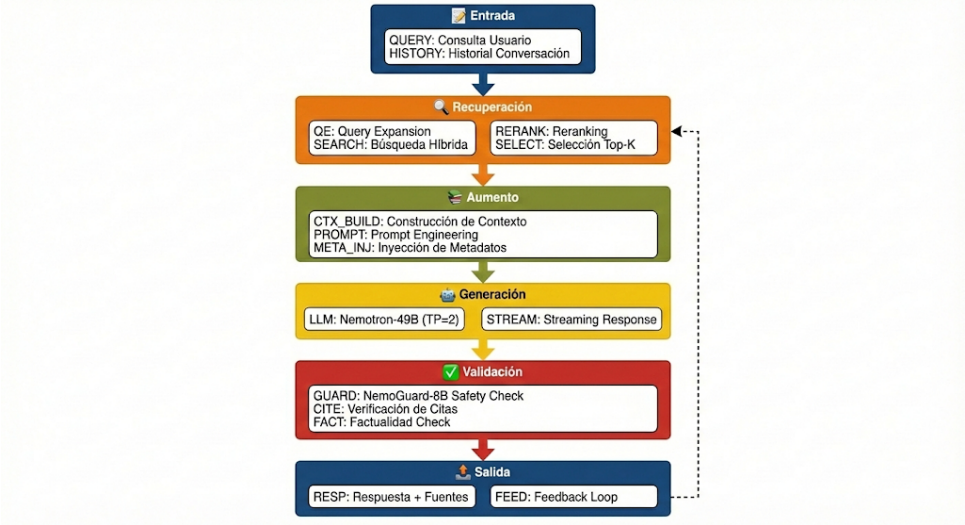
El modelo Nemotron-49B o similar, con su ventana de contexto de 110K tokens, puede procesar simultáneamente decenas de documentos recuperados, sintetizando información dispersa en respuestas estructuradas. Los guardrails implementados mediante NemoGuard-8B verifican la fidelidad de las respuestas al contexto recuperado, previniendo fabricación de información. Las siguientes secciones detallan cada componente de esta arquitectura generativa.

9.1 Arquitectura RAG del Agente BIEN

La arquitectura Retrieval-Augmented Generation (RAG) integra el motor de búsqueda semántico con el modelo de lenguaje grande para generar respuestas fundamentadas [REF001, REF002]. Como señala la investigación seminal de Lewis et al., "RAG models combine the benefits of pre-trained parametric and non-parametric memory for language generation" [REF001], permitiendo que el LLM produzca respuestas actualizadas y verificables sin necesidad de reentrenamiento. La implementación mediante NVIDIA RAG Blueprint [REF027, REF028] proporciona componentes optimizados para GPU que maximizan el rendimiento. El siguiente diagrama presenta la arquitectura completa.

Diagrama 9.1: Arquitectura RAG Completa del Agente BIEN

CONFIDENCIAL

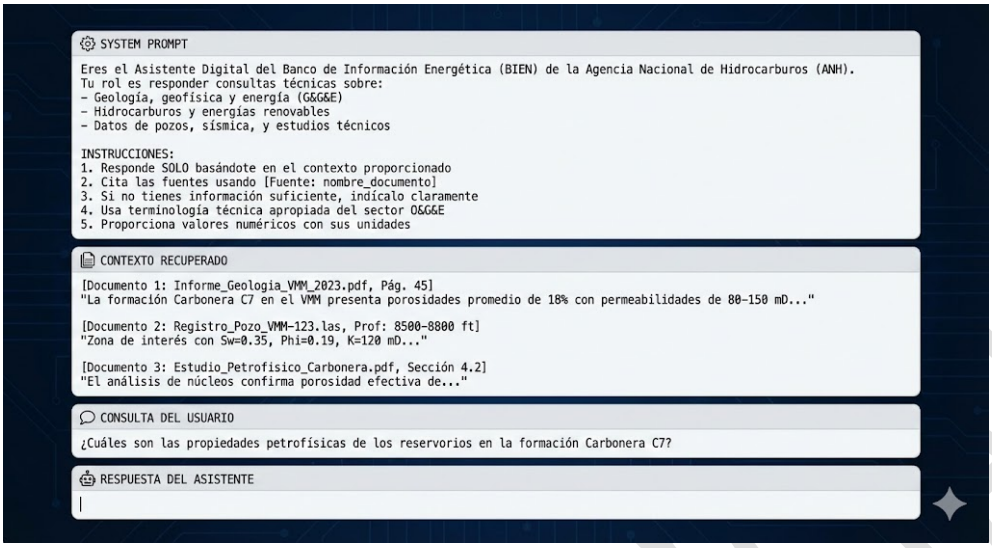


La arquitectura RAG implementa un ciclo de recuperación-aumento-generación-validación [REF002, REF028]. La fase de recuperación utiliza el motor de búsqueda semántico para identificar los documentos más relevantes. La fase de aumento construye un prompt estructurado que incluye el contexto recuperado, metadatos de las fuentes e instrucciones de comportamiento—estudios demuestran que "proper context formatting can improve answer quality by 15-30%" [REF027]. La fase de generación produce la respuesta mediante el LLM. La fase de validación verifica seguridad del contenido, precisión de las citas y consistencia factual con las fuentes. El feedback loop permite refinamiento iterativo si la respuesta inicial no satisface criterios de calidad.

9.2 Construcción del PROMPT RAG

La plantilla de PROMPT utilizada para el Agente BIEN sigue una estructura optimizada para el dominio de hidrocarburos.

Diagrama 9.2: Estructura del PROMPT RAG para el Agente BIEN



La estructura del prompt incluye: (1) System prompt que establece el rol, dominio y reglas de comportamiento del asistente; (2) Contexto recuperado con citas explícitas a documentos fuente, página/sección, y fragmentos relevantes; (3) Consulta del usuario tal como fue formulada; (4) Espacio para la respuesta generada. Esta estructura garantiza que el LLM tenga toda la información necesaria para producir respuestas precisas y verificables.

9.3 Parámetros de Generación del LLM

La siguiente tabla especifica los parámetros de generación configurados para el modelo Nemotron-49B en el contexto del BIEN.

Tabla 9.1: Configuración de Parámetros de Generación LLM

Parámetro	Valor	Justificación
temperature	0.3	Baja para respuestas técnicas consistentes
top_p	0.9	Nucleus sampling para diversidad controlada
top_k	50	Limita vocabulario a tokens más probables
max_tokens	2048	Suficiente para respuestas detalladas
repetition_penalty	1.1	Evita repetición de frases
presence_penalty	0.0	No penaliza temas ya mencionados
frequency_penalty	0.3	Reduce repetición de palabras específicas
stop_sequences	["[Fin]", ":"]	Delimita fin de respuesta
streaming	true	Respuesta progresiva para mejor UX

CONFIDENCIAL

La temperatura de 0.3 (baja) es crítica para respuestas técnicas donde la precisión factual prevalece sobre la creatividad. En contraste con chatbots creativos que usan temperaturas de 0.7-1.0, el Asistente BIEN debe producir respuestas consistentes y reproducibles. El streaming habilitado permite que el usuario comience a leer la respuesta mientras el modelo continúa generando, reduciendo la percepción de latencia.

10. Motores de Búsqueda Avanzada en la Plataforma BIEN

Esta sección describe la arquitectura, configuración y lineamientos técnicos para la implementación de los motores de búsqueda avanzados en la plataforma del BIEN, integrando las capacidades descritas en las secciones anteriores con el modelo de datos operacional del sistema.

10.1 Arquitectura de Integración

La arquitectura de motores de búsqueda avanzados en la plataforma BIEN integra capacidades de:

- Búsqueda estructurada: Consultas exactas sobre campos indexados
- Búsqueda semántica: Comprensión de significado mediante embeddings vectoriales
- Procesamiento de Lenguaje Natural (NLP): Análisis e interpretación de consultas
- Indexación documental inteligente: Extracción y catalogación automática de contenido

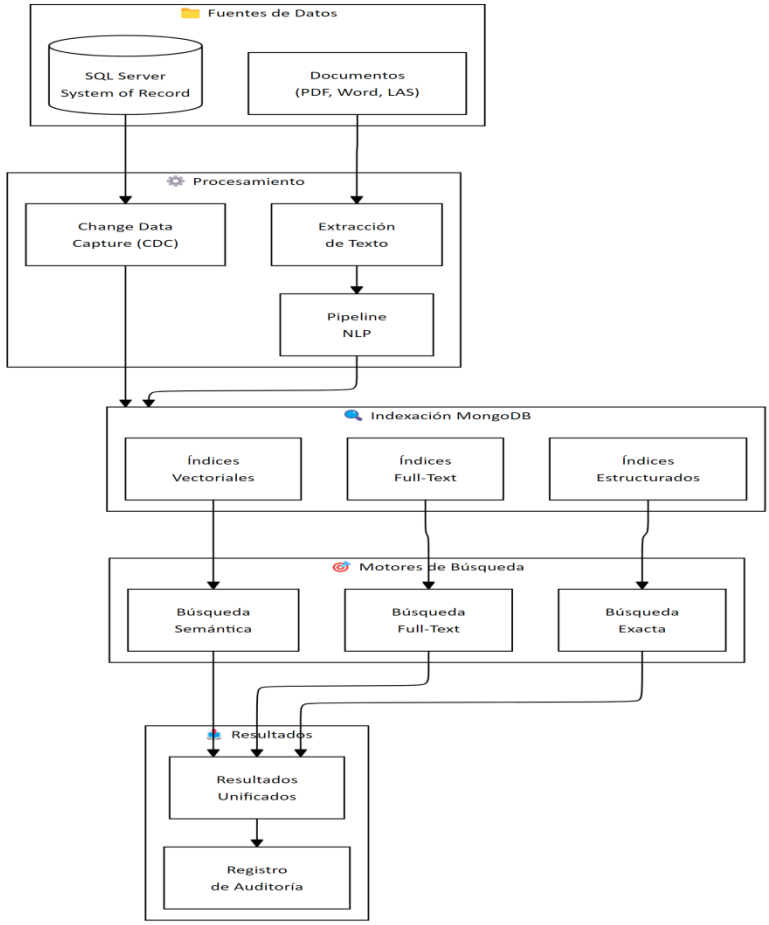
Todo ello alineado con el modelo de datos diseñado, donde:

- SQL Server actúa como *System of Record* (fuente autoritativa de datos)
- MongoDB funciona como *motor de indexación* para búsquedas avanzadas

La arquitectura es compatible con los estándares OSDU (Open Subsurface Data Universe) y DAMA-DMBOK para gestión de datos.

Diagrama 10.1: Arquitectura de Integración de Motores de Búsqueda en la Plataforma BIEN

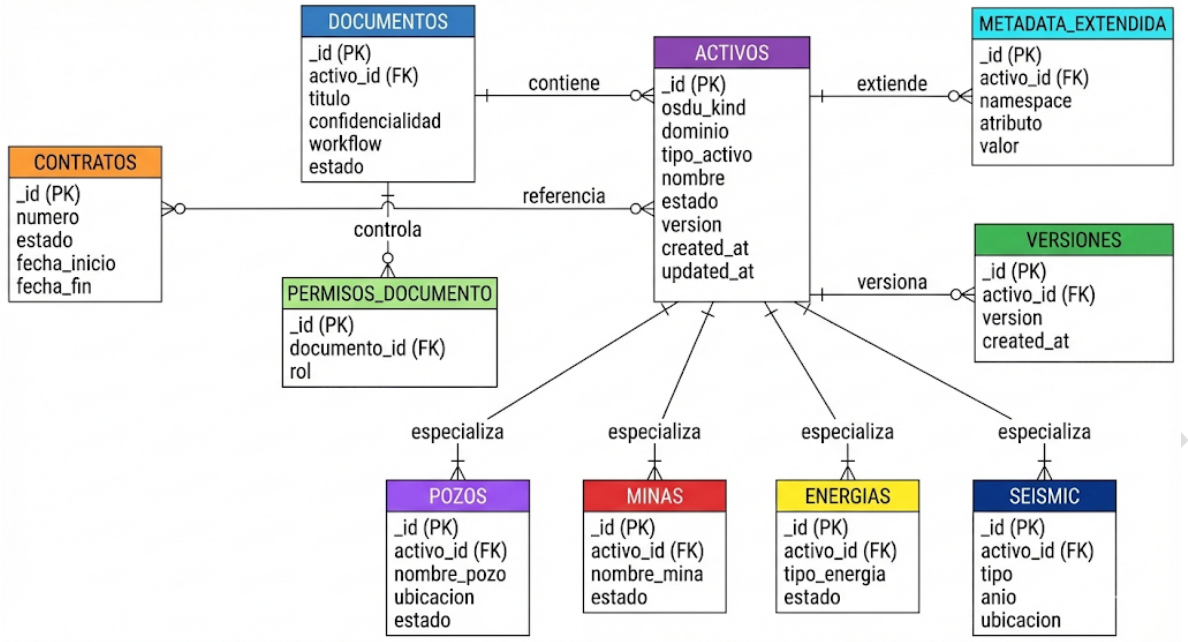
CONFIDENCIAL



10.2 Modelo de Indexación en MongoDB

Cada documento en MongoDB representa un ACTIVO_INFORMACION del BIEN, estructurado para soportar múltiples tipos de búsqueda simultáneamente.

Estructura del Documento de Índice



Esta estructura permite que un único documento soporte búsquedas estructuradas (por campos específicos), full-text (sobre texto indexado) y semánticas (mediante embeddings vectoriales) de manera simultánea.

10.3 Campos Utilizados para Búsqueda

Tabla 10.1: Campos para Búsqueda Estructurada (Exacta)

Campo	Uso	Ejemplo de Consulta
dominio	Filtro por área temática	dominio = "Pozos"
osdu_kind	Tipo OSDU del activo	osdu_kind = "osdu:wks:well:1.0.0"
estado	Estado del activo	estado = "ACTIVO"
clase_documental	Clasificación normativa	clase_documental = "Informe Técnico"
contrato.numero	Identificación contractual	contrato.numero = "491-2020"
metadata_clave.fecha_documento	Rango temporal	fecha_documento BETWEEN 2019 AND 2023
ubicacion.cuenca	Filtro geográfico	ubicacion.cuenca = "VMM"

Tabla 10.2: Campos para Búsqueda Textual (Full-Text)

Campo	Descripción	Índice MongoDB
nombre	Título del documento	Text index con peso 10
contrato.objeto	Objeto contractual	Text index con peso 5
texto_indexado	Texto completo extraído	Text index con peso 1

Tabla 10.3: Campos para Búsqueda Semántica

Campo	Descripción	Configuración
embedding	Vector semántico de 1024 dimensiones	Índice vectorial (cosine similarity)
texto_indexado	Contexto para NLP	Fuente para generación de embeddings

10.4 Configuración de Búsqueda Semántica

La búsqueda semántica en la plataforma BIEN permite:

- Consultas en lenguaje natural: Los usuarios pueden formular preguntas como lo harían a un experto
- Comprensión de intención: El sistema interpreta qué información busca el usuario
- Resultados por significado: Recuperación basada en similitud conceptual, no coincidencia literal

Ejemplo de Consulta Semántica:

"Contratos de supervisión financiera en pozos entre 2019 y 2021"

El sistema: 1. Extrae la intención (búsqueda de contratos) 2. Identifica entidades (supervisión financiera, pozos, rango temporal) 3. Genera embedding de la consulta 4. Busca documentos con embeddings similares 5. Aplica filtros estructurados (fecha, dominio) 6. Retorna resultados ordenados por relevancia semántica

10.5 Observabilidad y Control del Motor de Búsqueda

Toda interacción con los motores de búsqueda avanzada genera evidencia auditable, incluyendo:

CONFIDENCIAL

Tabla 10.4: Elementos de Auditoría por Consulta

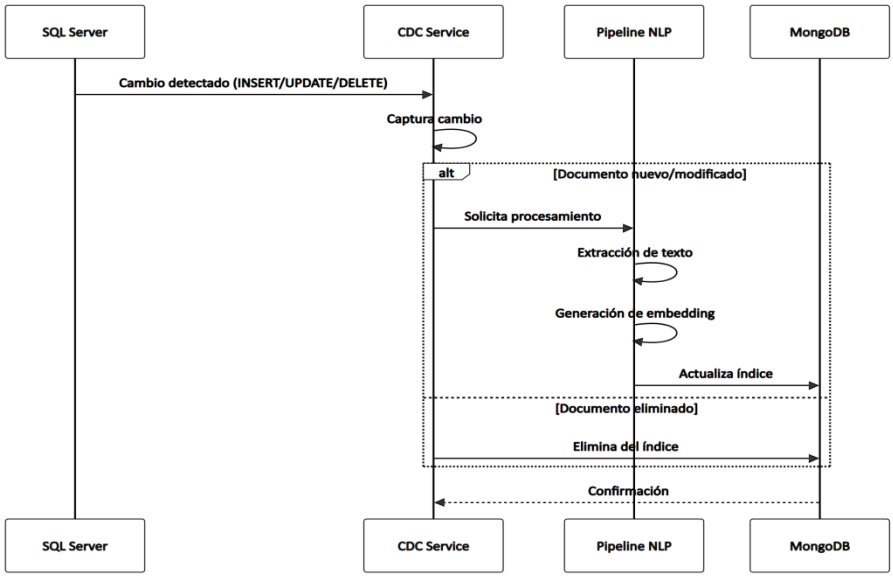
Elemento	Descripción	Uso
Consulta original	Texto exacto ingresado por el usuario	Análisis de patrones de uso
Índices utilizados	Estructurado, full-text, semántico o combinación	Optimización de rendimiento
Fuentes recuperadas	IDs de documentos retornados	Trazabilidad de información
Tiempo de respuesta	Latencia de la operación	Monitoreo de SLAs
Resultado autorizado	Documentos efectivamente mostrados tras filtros de seguridad	Cumplimiento de gobernanza

Este registro permite: - Monitoreo continuo del rendimiento del sistema - Mejora iterativa basada en patrones de consulta - Soporte a auditorías técnicas y funcionales - Análisis de uso para optimización de índices

10.6 Sincronización y Reprocesos

La sincronización entre SQL Server (fuente autoritativa) y MongoDB (motor de indexación) se gestiona mediante Change Data Capture (CDC):

Diagrama 10.2: Flujo de Sincronización CDC



Los cambios en SQL Server disparan automáticamente: - Reindexación de contenido textual - Actualización semántica (regeneración de embeddings) - Sincronización de metadatos estructurados

10.7 Beneficios del Diseño Integrado

La arquitectura de motores de búsqueda avanzados en la plataforma BIEN proporciona:

Beneficio	Descripción
Búsqueda rápida y flexible	Combinación de múltiples estrategias de búsqueda según necesidad
Comprensión semántica	Interpretación del significado de consultas en lenguaje natural
Gobernanza completa	Trazabilidad de accesos y auditoría de operaciones
Seguridad integrada	Filtrado de resultados según permisos del usuario
Reprocesamiento seguro	Sincronización automática ante cambios en fuentes autoritativas
Cumplimiento normativo	Alineación con estándares OSDU y DAMA-DMBOK

11. Conclusiones y Recomendaciones

11.1 Resumen de Capacidades Implementadas

La configuración de motores de búsqueda avanzados con capacidades semánticas y NLP para el Agente BIEN cumple integralmente con los requerimientos contractuales establecidos en el Anexo B del contrato ANH-002-CM-2025. Las capacidades implementadas incluyen:

1. Búsqueda Semántica Vectorial: Embeddings de 1024 dimensiones mediante NV-EmbedQA-1B, almacenados y consultados en Milvus con aceleración GPU (cuVS).
2. Procesamiento de Lenguaje Natural: Pipeline completo de NLP con tokenización, NER especializado para hidrocarburos, detección de intención y resolución de correferencia.
3. Búsqueda Híbrida: Combinación de búsqueda densa (semántica) y dispersa (full-text) con fusión RRF y reranking mediante NV-RerankQA-1B.
4. Generación Aumentada por Recuperación (RAG): Integración con LLM Nemotron-49B para síntesis de respuestas fundamentadas en el corpus documental.
5. Validación de Contenido: Guardrails mediante NemoGuard-8B para seguridad y verificación de citación de fuentes.

11.2 Alineación con Objetivos del BIEN

La siguiente tabla resume la alineación entre los objetivos estratégicos del BIEN y las capacidades del motor de búsqueda semántico.

CONFIDENCIAL

Tabla 11.1: Alineación Objetivos Estratégicos vs Capacidades

Objetivo BIEN	Capacidad del Motor	Beneficio
Centralización de información G&G&E	Indexación unificada de ~1.743,500 documentos	Búsqueda en todo el corpus desde interfaz única
Disponibilidad ágil de información	Latencia <1.5s para 95% de consultas	Respuestas casi instantáneas para analistas
Búsquedas inmersivas semánticas	NLP + Embeddings + RAG	Consultas en lenguaje natural con comprensión de significado
Identificación de insights	LLM con capacidad de síntesis	Generación automática de resúmenes y análisis
Soporte a toma de decisiones	Respuestas fundamentadas con citación	Información verificable y trazable

11.3 Recomendaciones para Optimización Continua

1. Fine-tuning del modelo de embedding: Entrenar NV-EmbedQA-1B con corpus específico del BIEN para mejorar representación de terminología técnica de hidrocarburos colombianos.
2. Expansión del vocabulario NER: Actualizar continuamente el diccionario de entidades con nuevos nombres de pozos, bloques y proyectos.
3. Evaluación periódica RAGAS: Ejecutar benchmarks mensuales con conjunto de prueba curado para monitorear calidad del sistema.
4. Optimización de índices: Revisar configuración de índices Milvus trimestralmente conforme crece el corpus.
5. Feedback loop de usuarios: Implementar mecanismo de retroalimentación para identificar consultas problemáticas y mejorar iterativamente.

Referencias

Las referencias bibliográficas completas utilizadas en este informe se encuentran documentadas en fuentes principales organizadas por categoría, anexo y demás que continúan en el presente documento:

Fuentes Primarias

Ref	Documento	Descripción
-	Anexo B Contrato ANH-002-CM-2025	Requisitos contractuales del BIEN

CONFIDENCIAL

Código	Título	Relevancia
REF001	Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (Lewis et al., 2020)	Fundamentos teóricos RAG
REF002	REALM: Retrieval-Augmented Language Model Pre-Training (Guu et al., 2020)	Arquitectura RAG
REF003	A Survey on Retrieval-Augmented Text Generation (Li et al., 2022)	Estado del arte RAG
REF006	Dense Passage Retrieval for Open-Domain QA (Karpukhin et al., 2020)	Embeddings densos
REF007	Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks	Embeddings semánticos
REF009	Embeddings in Machine Learning (Google ML Guide, 2024)	Fundamentos embeddings
REF012	Semantic Chunking for RAG Systems (LlamaIndex, 2024)	Estrategias de chunking
REF013	Evaluating Chunking Strategies for RAG (LangChain, 2024)	Evaluación de chunking
REF016	Hybrid Search Combining Dense and Sparse Retrieval (Ma et al., 2022)	Búsqueda híbrida
REF017	Pinecone Hybrid Search: Sparse-Dense Embeddings (2024)	Implementación híbrida
REF018	Reciprocal Rank Fusion Outperforms Condorcet (Cormack et al., 2009)	Algoritmo RRF
REF019	Cross-Encoder vs Bi-Encoder for Semantic Search	Reranking
REF020	Hybrid Search Experiments: When to Use What (Weaviate, 2024)	Benchmarks híbridos
REF022	Speech and Language Processing (Jurafsky & Martin, 2023)	Fundamentos NLP
REF025	Named Entity Recognition with BERT (Devlin et al., 2019)	NER con Transformers
REF027	NVIDIA RAG Blueprint Documentation v2.3.0	Implementación NVIDIA
REF028	NeMo Retriever Documentation	Componentes NIM
REF029	NV-EmbedQA Model Card (NVIDIA, 2024)	Modelo de embeddings
REF039	RAGAS: Automated Evaluation of RAG Pipelines (Es et al., 2023)	Framework evaluación
REF040	RAGAS Documentation - Metrics Guide	Métricas RAGAS
REF041	Evaluating RAG Systems in Production (Microsoft, 2024)	Evaluación producción

CONFIDENCIAL

Recursos Técnicos

- NVIDIA RAG Blueprint Documentation v2.3.0 - https://docs.nvidia.com/rag/
- Milvus Vector Database Documentation - https://milvus.io/docs
- RAGAS Framework Documentation - https://docs.ragas.io/
- LangChain RAG Documentation - https://python.langchain.com/docs/tutorials/rag/
- LlamaIndex RAG Guide - https://docs.llamaindex.ai/

Documento elaborado para el Proyecto BIEN - ANH Colombia Configuración de Motores de Búsqueda Semánticos y NLP Versión 1.0 - Enero 2026

CONFIDENCIAL

Anexo - Tabla de Definiciones

La siguiente tabla consolida las definiciones técnicas de los términos utilizados en el informe de configuración de motores de búsqueda semánticos y NLP para el Agente BIEN. Los términos están organizados alfabéticamente para facilitar la consulta.

Término	Definición	Contexto de Uso en BIEN
Alucinación (Hallucination)	Fenómeno donde un modelo de lenguaje grande genera información falsa, inventada o no fundamentada en los datos de entrada. Las alucinaciones ocurren cuando el modelo "inventa" hechos que suenan plausibles pero no están respaldados por evidencia.	El sistema RAG del BIEN utiliza fundamentación en documentos recuperados y guardrails (NemoGuard-8B) para reducir alucinaciones en las respuestas del Asistente Digital.
Batch Processing	Técnica de procesamiento que agrupa múltiples elementos (textos, consultas, embeddings) en lotes para procesamiento simultáneo, maximizando la utilización de recursos GPU y reduciendo latencia por elemento.	La ingesta de documentos del BIEN procesa chunks en batches de 32-64 elementos para optimizar el throughput de generación de embeddings.
BM25 (Best Matching 25)	Algoritmo de ranking probabilístico para recuperación de información basado en frecuencia de términos (TF) y frecuencia inversa de documentos (IDF). Considera la longitud del documento y la saturación de términos. Es el estándar de facto para búsqueda léxica.	El BIEN utiliza BM25 como componente de búsqueda dispersa (sparse) en el sistema híbrido, complementando la búsqueda densa para capturar coincidencias exactas de términos técnicos.
Chunking	Proceso de segmentación de documentos extensos en fragmentos más pequeños (chunks) adecuados para indexación y recuperación. La estrategia de chunking afecta directamente la calidad de la búsqueda semántica.	Los documentos del BIEN se segmentan en chunks de 500-1000 tokens con 10-20% de overlap para preservar contexto semántico en los informes técnicos.
Contexto (Context Window)	Cantidad máxima de tokens que un modelo de lenguaje puede procesar simultáneamente en una sola	El modelo Nemotron-49B del BIEN soporta 131K tokens de contexto, permitiendo incluir múltiples documentos técnicos extensos en una sola consulta.

CONFIDENCIAL

	inferencia. Determina cuánta información recuperada puede incluirse en el prompt del LLM.	
Correferencia (Coreference Resolution)	Tarea de NLP que identifica y vincula todas las menciones en un texto que refieren a la misma entidad (ej: "el pozo", "esta perforación", "Cusiana-1" → misma entidad).	Permite que el sistema BIEN comprenda referencias pronominales en documentos técnicos extensos donde se menciona un pozo múltiples veces con diferentes expresiones.
Coseno, Similitud	Métrica que mide el ángulo entre dos vectores en el espacio de embeddings. Valores cercanos a 1 indican alta similitud semántica; valores cercanos a 0 indican baja similitud. No considera la magnitud de los vectores, solo su dirección.	El BIEN utiliza similitud coseno como métrica principal para comparar embeddings de consultas contra embeddings de documentos indexados en Milvus.
Cross-Encoder	Arquitectura de modelo que procesa conjuntamente el par (consulta, documento) para producir un score de relevancia. Más preciso que bi-encoders pero computacionalmente más costoso.	NV-RerankQA-1B actúa como cross-encoder en el BIEN, reordenando los top-100 candidatos de la búsqueda inicial para mejorar precisión.
cuVS (CUDA Vector Search)	Librería de NVIDIA para búsqueda de vectores acelerada por GPU. Implementa algoritmos de búsqueda aproximada de vecinos más cercanos (ANN) optimizados para GPUs NVIDIA.	Milvus integra cuVS para acelerar las búsquedas vectoriales del BIEN en las GPUs H100, logrando latencias de búsqueda inferiores a 50ms.
Dense Retrieval	Técnica de recuperación que representa consultas y documentos como vectores densos (embeddings) y recupera documentos por proximidad en el espacio vectorial. Captura similitud semántica.	Componente principal de búsqueda del BIEN, utilizando NV-EmbedQA-1B para generar embeddings de 1024 dimensiones.
Embedding	Representación numérica de datos (texto, imagen, audio) en un espacio vectorial de alta dimensionalidad donde la proximidad geométrica refleja similitud semántica. Los embeddings son el	Cada chunk de documento del BIEN se transforma en un embedding de 1024 dimensiones que captura su significado semántico para búsqueda por similitud.

CONFIDENCIAL

	fundamento de la búsqueda semántica moderna.	
Euclidiana, Distancia	Métrica que mide la distancia en línea recta entre dos puntos (vectores) en el espacio. Menor distancia indica mayor similitud. Considera tanto dirección como magnitud de los vectores.	Alternativa a similitud coseno disponible en Milvus, aunque coseno es preferido para embeddings de texto normalizados.
Faithfulness	Métrica de evaluación RAG que mide qué proporción de las afirmaciones en la respuesta generada están respaldadas por el contexto recuperado. Indica el grado de fundamentación de la respuesta.	Objetivo del BIEN: Faithfulness > 0.90, asegurando que las respuestas del Asistente Digital estén fundamentadas en los documentos técnicos.
Fine-tuning	Proceso de adaptar un modelo pre-entrenado a un dominio o tarea específica mediante entrenamiento adicional con datos del dominio objetivo.	Recomendación para el BIEN: fine-tuning del modelo NER con corpus de hidrocarburos colombianos para mejorar reconocimiento de entidades del dominio.
Full-Text Search	Búsqueda que indexa y recupera documentos basándose en el contenido textual completo, típicamente usando índices invertidos y algoritmos como BM25 o TF-IDF.	Componente de Elasticsearch en el BIEN para búsqueda por palabras clave exactas, complementando la búsqueda semántica.
GPU (Graphics Processing Unit)	Procesador especializado diseñado originalmente para gráficos, ahora ampliamente utilizado para computación paralela en inteligencia artificial debido a su arquitectura de múltiples núcleos.	El servidor BIEN cuenta con 4× NVIDIA H100 NVL (376 GB VRAM total) para ejecutar los modelos de IA del sistema RAG.
Guardrails	Mecanismos de seguridad y control que limitan el comportamiento de modelos de IA para evitar respuestas dañinas, inapropiadas o fuera del dominio autorizado.	NemoGuard-8B implementa guardrails en el BIEN para validar que las respuestas sean seguras y apropiadas para el contexto institucional de la ANH.
HBM3 (High Bandwidth Memory 3)	Tipo de memoria de alto ancho de banda utilizada en GPUs de alto rendimiento. Proporciona mayor velocidad de transferencia de datos que	Las GPUs H100 NVL del BIEN utilizan 94 GB de HBM3 cada una, proporcionando el ancho de banda necesario para modelos LLM de gran escala.

CONFIDENCIAL

	la memoria GDDR convencional.	
HNSW (Hierarchical Navigable Small World)	Algoritmo de búsqueda aproximada de vecinos más cercanos que construye un grafo jerárquico navegable para búsquedas eficientes en espacios de alta dimensionalidad.	Tipo de índice disponible en Milvus para el BIEN, ofreciendo balance entre precisión y velocidad de búsqueda.
Hybrid Search	Estrategia de búsqueda que combina múltiples métodos de recuperación (típicamente dense + sparse) para maximizar tanto precisión semántica como coincidencia léxica exacta.	El BIEN implementa búsqueda híbrida combinando embeddings densos (Milvus) con BM25 (Elasticsearch), fusionados mediante RRF.
Indexación	Proceso de organizar y estructurar datos para permitir búsquedas eficientes. En bases de datos vectoriales, incluye la creación de estructuras como IVF, HNSW o FLAT para búsqueda por similitud.	Los ~7-10 millones de chunks del BIEN se indexan en Milvus con índices IVF_FLAT para búsqueda vectorial de baja latencia.
Inferencia	Proceso de ejecutar un modelo de IA entrenado para generar predicciones o salidas a partir de nuevas entradas. En LLMs, es la generación de texto basada en un prompt.	El servidor BIEN ejecuta inferencia de Nemotron-49B en GPUs H100 para generar respuestas a consultas de usuarios.
Intención (Intent Detection)	Tarea de NLP que clasifica el propósito o objetivo subyacente de una consulta de usuario (ej: búsqueda factual, comparación, síntesis, listado).	El BIEN clasifica consultas en categorías de intención para optimizar la estrategia de búsqueda y generación de respuestas.
IVF_FLAT	Tipo de índice vectorial que particiona el espacio de embeddings en clusters (Inverted File) y realiza búsqueda exhaustiva (FLAT) dentro de los clusters seleccionados. Balance entre precisión y eficiencia.	Índice predeterminado del BIEN en Milvus: nlist=4096 clusters, nprobe=128 clusters explorados por consulta.
k-NN (k-Nearest Neighbors)	Algoritmo que encuentra los k elementos más cercanos a un punto de consulta en un espacio métrico. Base de la	Las consultas del BIEN recuperan los k=100 chunks más cercanos semánticamente antes del reranking.

CONFIDENCIAL

	búsqueda vectorial por similitud.	
Latencia	Tiempo transcurrido desde que se envía una consulta hasta que se recibe la respuesta completa. Métrica crítica de rendimiento para sistemas interactivos.	Objetivos del BIEN: P50 < 500ms, P95 < 1.5s para consultas del Asistente Digital.
LLM (Large Language Model)	Modelo de inteligencia artificial basado en arquitectura transformer, entrenado en grandes corpus de texto, capaz de comprender y generar lenguaje natural con alta fluidez.	Nemotron-49B es el LLM del BIEN para generación de respuestas contextualizadas basadas en documentos recuperados.
Milvus	Base de datos vectorial open-source diseñada para almacenamiento y búsqueda de embeddings a escala. Soporta múltiples tipos de índices y aceleración GPU mediante cuVS.	Almacén vectorial principal del BIEN para los ~7-10 millones de embeddings de chunks documentales.
Multimodal	Sistemas de IA capaces de procesar y relacionar múltiples tipos de datos (texto, imagen, audio, video) en un espacio de representación unificado.	El VLM Nemotron-12B-VL del BIEN procesa imágenes técnicas (mapas sísmicos, registros de pozos) generando descripciones textuales indexables.
NER (Named Entity Recognition)	Tarea de NLP que identifica y clasifica menciones de entidades nombradas (personas, organizaciones, lugares, fechas, etc.) en texto no estructurado.	El BIEN utiliza NER especializado para extraer entidades del dominio de hidrocarburos: pozos, cuencas, formaciones, coordenadas.
NIM (NVIDIA Inference Microservices)	Microservicios de inferencia pre-optimizados de NVIDIA, empaquetados como contenedores Docker para despliegue simplificado de modelos de IA en infraestructura NVIDIA.	Los componentes de IA del BIEN (embedding, reranking, LLM) se despliegan como NIMs para gestión y escalabilidad simplificadas.
NLP (Natural Language Processing)	Campo de la inteligencia artificial dedicado a la interacción entre computadoras y lenguaje humano, incluyendo comprensión, generación y traducción de texto.	El BIEN integra múltiples componentes NLP: tokenización, NER, detección de intención, embeddings y generación de texto.

CONFIDENCIAL

nprobe	Parámetro de búsqueda en índices IVF que determina cuántos clusters se exploran durante una consulta. Mayor nprobe = mayor precisión pero mayor latencia.	Configuración del BIEN: nprobe=128 de 4096 clusters (~3%), balance entre precisión (>99%) y velocidad (<50ms).
NUMA (Non-Uniform Memory Access)	Arquitectura de memoria en servidores multi-socket donde el tiempo de acceso a memoria depende de la ubicación relativa del procesador y la memoria.	El servidor BIEN tiene 2 nodos NUMA: GPUs 0-1 en NUMA0 (LLM), GPUs 2-3 en NUMA1 (recuperación), optimizando localidad de memoria.
NVLink	Tecnología de interconexión de alta velocidad de NVIDIA para comunicación entre GPUs. Proporciona ancho de banda muy superior a PCIe para cargas de trabajo distribuidas.	Las GPUs H100 del BIEN están conectadas por NVLink (900 GB/s), habilitando Tensor Parallelism eficiente para el LLM.
OCR (Optical Character Recognition)	Tecnología que convierte imágenes de texto (documentos escaneados, fotografías) en texto digital editable y procesable por NLP.	PaddleOCR/NeMo extrae texto de los ~173,000 PDFs escaneados del BIEN para indexación y búsqueda semántica.
Ontología	Representación formal del conocimiento de un dominio, incluyendo conceptos, categorías, propiedades y relaciones entre entidades.	La ontología semántica del BIEN organiza el conocimiento de hidrocarburos (cuencas, formaciones, pozos, recursos) para búsqueda estructurada.
Overlap	Cantidad de tokens compartidos entre chunks consecutivos de un documento segmentado. Preserva contexto en las fronteras de fragmentación.	Configuración del BIEN: 10-20% overlap (100-200 tokens) para chunks de 1000 tokens, evitando pérdida de información en fronteras.
Pipeline	Secuencia de etapas de procesamiento donde la salida de una etapa alimenta la entrada de la siguiente. Permite modularidad y optimización de flujos de datos.	El pipeline RAG del BIEN: consulta → NLP → embedding → búsqueda → reranking → generación → validación → respuesta.
POS Tagging (Part-of-Speech)	Tarea de NLP que asigna categorías gramaticales (sustantivo, verbo, adjetivo, etc.) a cada palabra de una oración, facilitando análisis sintáctico posterior.	Componente del pipeline NLP del BIEN para desambiguar términos técnicos polisémicos del dominio de hidrocarburos.

CONFIDENCIAL

Precision	Métrica de evaluación que mide la proporción de resultados relevantes entre todos los resultados recuperados. Precision = Relevantes_Recuperados / Total_Recuperados.	Objetivo del BIEN: Context Precision > 0.80, asegurando que los chunks recuperados sean relevantes para la consulta.
Prompt	Texto de entrada proporcionado a un LLM que incluye instrucciones, contexto y/o ejemplos para guiar la generación de la respuesta deseada.	El prompt RAG del BIEN incluye: system prompt (rol/reglas), contexto recuperado (chunks), y consulta del usuario.
Query Expansion	Técnica que enriquece la consulta original con términos relacionados, sinónimos o contexto adicional para mejorar la recuperación de información.	El pipeline del BIEN puede expandir "pozos exploratorios" a incluir "wildcat wells", "perforaciones de exploración" para búsqueda multilingüe.
RAG (Retrieval-Augmented Generation)	Arquitectura de IA que combina recuperación de información con generación de lenguaje: primero recupera documentos relevantes, luego los utiliza como contexto para que un LLM genere respuestas fundamentadas.	Arquitectura central del Agente BIEN, utilizando NVIDIA RAG Blueprint para respuestas técnicas fundamentadas en el corpus documental.
RAGAS	Framework de evaluación para sistemas RAG que mide métricas como Faithfulness, Answer Relevancy, Context Precision y Context Recall de forma automatizada.	El BIEN utiliza RAGAS para evaluación continua de la calidad del sistema RAG con objetivo de >85% en métricas principales.
Ranking	Proceso de ordenar resultados de búsqueda por relevancia respecto a la consulta. Incluye ranking inicial (retrieval) y re-ranking (refinamiento).	El BIEN implementa ranking en dos etapas: búsqueda híbrida inicial (top-100) + cross-encoder reranking (top-10 final).
Recall	Métrica de evaluación que mide la proporción de documentos relevantes recuperados respecto al total de documentos relevantes existentes. Recall = Relevantes_Recuperados / Total_Relevantes.	Objetivo del BIEN: Context Recall > 0.85, asegurando que la mayoría de información necesaria se recupere.

CONFIDENCIAL

Reranking	Proceso de reordenar un conjunto inicial de candidatos recuperados utilizando un modelo más sofisticado (típicamente cross-encoder) para mejorar la precisión del ranking final.	NV-RerankQA-1B del BIEN reordena los top-100 candidatos de la búsqueda híbrida para producir los top-10 más relevantes.
RRF (Reciprocal Rank Fusion)	Algoritmo de fusión de rankings que combina múltiples listas ordenadas ponderando documentos inversamente proporcionales a su posición en cada lista. Robusto y no requiere calibración de scores.	El BIEN fusiona resultados de búsqueda densa (Milvus) y dispersa (Elasticsearch) mediante RRF antes del reranking.
Semantic Search	Búsqueda que comprende el significado de la consulta y los documentos, encontrando resultados conceptualmente relevantes aunque no contengan las palabras exactas de la consulta.	Capacidad central del BIEN: permite que "pozos exploratorios Llanos" encuentre documentos sobre "wildcat wells Eastern Llanos Basin".
Sparse Retrieval	Técnica de recuperación basada en representaciones dispersas (la mayoría de valores son cero), típicamente usando frecuencia de términos como BM25 o TF-IDF.	Componente de Elasticsearch en el BIEN para capturar coincidencias exactas de términos técnicos únicos (nombres de pozos, códigos).
Streaming	Técnica de transmisión de respuestas de LLM token por token a medida que se generan, en lugar de esperar a la respuesta completa. Mejora percepción de latencia.	El Asistente Digital del BIEN utiliza streaming para que usuarios vean respuestas progresivamente mientras el LLM genera.
Temperature	Parámetro de generación de LLM que controla la aleatoriedad de las predicciones. Valores bajos producen salidas más determinísticas; valores altos, más creativos/diversos.	El BIEN usa temperature=0.3 (baja) para respuestas técnicas consistentes y reproducibles, minimizando variabilidad.
Tensor Parallelism (TP)	Técnica de paralelización que distribuye los tensores (pesos) de un modelo entre múltiples GPUs, permitiendo ejecutar modelos que	Nemotron-49B del BIEN se ejecuta con TP=2 en GPUs 0+1, dividiendo el modelo entre ambas GPUs conectadas por NVLink.

CONFIDENCIAL

	exceden la memoria de una sola GPU.	
Throughput	Cantidad de operaciones completadas por unidad de tiempo. En sistemas RAG, típicamente medido en consultas por minuto o documentos procesados por hora.	Objetivos del BIEN: >100 queries/min para consultas, >2000 docs/hora para ingesta de nuevos documentos.
Tokenización	Proceso de dividir texto en unidades discretas (tokens) que el modelo puede procesar. Los tokens pueden ser palabras, subpalabras, caracteres o combinaciones, dependiendo del tokenizador.	El BIEN usa SentencePiece/BPE para tokenización, manejando vocabulario técnico de hidrocarburos y términos en español e inglés.
Top-k	Parámetro que especifica el número de resultados/candidatos a recuperar o generar. En búsqueda, limita resultados; en generación, limita tokens considerados.	El BIEN recupera top-k=100 candidatos iniciales, reranked a top-k=10 finales para contexto del LLM.
Transformer	Arquitectura de red neuronal basada en mecanismos de atención (attention), fundamento de los modelos de lenguaje modernos como GPT, BERT, Llama.	Todos los modelos del BIEN (Nemotron-49B, NV-EmbedQA-1B, NV-RerankQA-1B) están basados en arquitectura transformer.
Vector Database	Base de datos especializada en almacenar, indexar y buscar vectores de alta dimensionalidad (embeddings), optimizada para búsquedas de similitud a escala.	Milvus es la base de datos vectorial del BIEN, almacenando ~7-10 millones de embeddings de 1024 dimensiones.
VLM (Vision-Language Model)	Modelo de IA capaz de procesar conjuntamente imágenes y texto, típicamente generando descripciones textuales de imágenes o respondiendo preguntas sobre contenido visual.	Nemotron-12B-VL del BIEN genera captions de imágenes técnicas (mapas sísmicos, registros) para indexación en el sistema de búsqueda.
VRAM (Video RAM)	Memoria dedicada de la GPU para almacenamiento de datos de procesamiento gráfico y computación. En IA, determina el tamaño máximo	El BIEN dispone de 376 GB de VRAM total (4×94 GB H100 NVL), suficiente para todos los componentes del sistema RAG.

CONFIDENCIAL

	del modelo y batch que puede procesarse.	

CONFIDENCIAL