ANH
AGENCIA NACIONAL DE HIDROCARBUROS

Contratar el análisis, diseño, implementación y puesta en funcionamiento del Banco de Información Energética (BIEN(Banco de Información Energética)) 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).



Infraestructura de datos estructurados y no estructurados

Bogotá, Colombia

DOCUMENTO CONFIDENCIAL

UT ANH PWD – NX 2025

Historial de versiones

password NEXURA

Unión Temporal ANH PWD-NX 2025

Id	Fecha (dd/mm/aa)	Versión	Descripción	Elaborado por	Revisado por
1	15/09/2025	1.0	Versión inicial del documento	UT ANH PWD-NX 2025	Director del Proyecto

CONFIDENCIAL

Unión Temporal ANH PWD-NX 2025

CONTENIDO

1. INTRODUCCIÓN Y PROPÓSITO 4
2. ALCANCE 4
3. RESUMEN EJECUTIVO 4
4. REFERENCIAS DE PROYECTO 5
5. ARQUITECTURA OBJETIVO 5
5.1 Datos Estructurados (SQL Server) 5
5.2 Datos No Estructurados / Semi-estructurados (MongoDB y Documental) 6
5.3 Integración GIS 6
5.4 Catálogo, Metadatos y Linaje 6
5.5 Seguridad e Identidad 7
5.6 Modelos de Datos (Referencia) 7
5.7 Stack Tecnológico y Despliegue Físico 7
5.8 Integraciones Institucionales y Contratos API 8
6. COMPONENTES Y CAPACIDADES 9
6.1 Inventario de Tablas (SQL Server) 9
6.2 Inventario de Colecciones (MongoDB) 9
6.3 Zonas y Esquema de Almacenamiento 9
6.4 Ingesta y Procesamiento 9
6.5 Exposición y Consumo 10
6.6 Parámetros de Servicio (SLI/SLO) y Capacidad 10
6.7 Alineación con Modelos ER y Diccionario de Datos 10
7. OPERACIÓN Y GOBIERNO 11
7.1 Gobierno de Datos 11
7.2 Seguridad y Cumplimiento 11
7.3 Respaldo y Recuperación 11
7.4 Monitoreo y Observabilidad 12
7.5 Matriz RACI 12
7.6 Plan de Continuidad (Backups/Restore/DRP) 14
8. MAPA DE REQUERIMIENTOS - CAPACIDADES 14
9. CONSIDERACIONES DE DESPLIEGUE 15
9.1 RTM: Requerimientos - Capacidades/Secciones 15
10. ANEXOS 15

CONFIDENCIAL

1. Introducción y Propósito

Este documento define la arquitectura y lineamientos de infraestructura para el manejo de datos estructurados y no estructurados del proyecto BIEN (ANH). Se priorizan las últimas versiones de los artefactos técnicos compartidos en este compendio —incluyendo modelos de datos (SQL Server y MongoDB), requerimientos funcionales (RF-01 a RF-10), requerimientos no funcionales y técnicos, guías de implementación, patrones de integración y reportes de arquitectura— con el fin de establecer un diseño ejecutable, gobernable y seguro.

2. Alcance

• Almacenamiento y procesamiento para datos estructurados (SQL Server) y no estructurados/semi-estructurados (MongoDB y repositorio documental).

• Integración con módulos clave: RF-02 Gestión Documental, RF-03 Geovisor GIS, RF-04 Analítica y Reportes, RF-05 Asistente de IA On-Prem, RF-06 Catálogo/Metadatos/Linaje, RF-07 Integraciones y Workflow, RF-08 Identidad/Acceso/Auditoría, RF-09 Ingesta continua/operación evolutiva, RF-10 Organización de Rondas y Licitaciones.

• Servicios transversales de seguridad, gobernanza, monitoreo/observabilidad, respaldo/recuperación (BCP/DR) y operación (Mesa de Ayuda).

3. Resumen Ejecutivo

El documento INFRAESTRUCTURA PARA DATOS ESTRUCTURADOS / NO ESTRUCTURADOS consolida la arquitectura y la del BIEN, alineada con el Anexo Técnico y los pliegos del contrato. La solución contempla SQL Server para datos transaccionales y maestros; MongoDB y un repositorio documental para datos no estructurados/semi-estructurados; y capacidades transversales de seguridad, auditoría, observabilidad y continuidad del negocio.

Se priorizaron las versiones finales de los modelos ER (SQL Server y MongoDB), los requerimientos funcionales RF-01 a RF-10, los requerimientos no funcionales/técnicos, y las decisiones arquitectónicas registradas. Se establecen inventarios de tablas/colecciones, validadores de esquema, prácticas de ingesta/curación/serving y controles de gobierno del dato.

CONFIDENCIAL

El diseño satisface las fases y entregables contractuales: levantamiento, arquitectura, implementación por dominios, pruebas, UAT, go-live controlado y estabilización. La arquitectura por capas (presentación, servicios, orquestación, datos, IA/ML, GIS y observabilidad) garantiza trazabilidad extremo a extremo, disponibilidad y seguridad.

• Cumplimiento: identidad federada con SSO/MFA; auditoría unificada (audit_logs); observabilidad E2E; RTM y evidencias de QA.

• Gobernanza: catálogo de metadatos, linaje, clasificación/confidencialidad y políticas de retención.

• Continuidad: respaldo/recuperación (RPO/RTO por dominio) y ambientes DEV/QA/PROD con CI/CD.

• GIS & Sísmica: catálogo de productos/capas y entidades sísmicas (línea 2D, stack 2D, volumen 3D, procesamiento y modelos de velocidad).

4. Referencias de Proyecto

• Modelos ER (SQL Server y MongoDB) finales en Mermaid (.mmd).

• Requerimientos Funcionales RF-01 … RF-10.

• Requerimientos No Funcionales / Técnicos (RNF/RT).

• Diagrama de Arquitectura de Software (ARQ_SAD_BIEN).

• Guía de Implementación y Patrones de Integración (.md).

• Contrato / Anexo Técnico y Manual de Identidad Visual.

5. Arquitectura Objetivo

5.1 Datos Estructurados (SQL Server)

El almacenamiento para datos estructurados se basa en SQL Server. Se definen esquemas y tablas para dominios clave —contratos, áreas contractuales, GIS (catálogo de productos y capas), sísmica (línea 2D, stack 2D, volumen 3D, procesamiento y modelos de velocidad), y pozos (pozo/ramal/registros)—. La normalización y la auditoría se implementan mediante convenciones snake_case y la tabla transversal audit_logs.

CONFIDENCIAL

• Esquema de datos consolidado (ver 5.1 Inventario de Tablas).

• Auditoría unificada: audit_logs con created_by, created_at, updated_by, updated_at, status.

• Optimización: índices por identificadores (*_id), fechas (*_date) y campos de búsqueda (título/nombre/código).

5.2 Datos No Estructurados / Semi-estructurados (MongoDB y Documental)

MongoDB funciona como data lake operacional por dominio (documental, GIS, sísmica, pozos), con validación de esquema (JSON Schema) y convenciones snake_case. Los binarios (archivos) referenciados por dataset_file, administrative_document y otras entidades se gestionan en repositorio documental con metadatos, checksum e integraciones de ingesta.

• Colecciones normalizadas (ver 5.2 Inventario de Colecciones) con validadores JSON.

• Referencias a archivos mediante file_refs/location/checksum.

• Metadatos y linaje integrados a Catálogo/Linaje (RF-06).

5.3 Integración GIS

La cartografía se orquesta con map_product, map_layer_vector y map_layer_raster, soportadas en crs. Las capas se indexan por bbox, crs y atributos para exposición a servicios de mapas y búsquedas espaciales.

5.4 Catálogo, Metadatos y Linaje

Se mantiene catálogo de datasets, colecciones y artefactos (p. ej., dataset_file, document_set, technical_report). El linaje y gobierno se reflejan en campos legal_tags, acl y ancestry; en SQL se centraliza auditoría con audit_logs. El módulo RF-06 concentra la consulta y curaduría.

CONFIDENCIAL

5.5 Seguridad e Identidad

• Control de acceso por roles (RF-08): administrador de datos, oficial de seguridad, auditor SIEM, operadores (GIS/BI/Documental).

• Auditoría completa (creación/actualización/estado) y consolidación en audit_logs.

• Cifrado en tránsito (TLS) y manejo de clasificación/confidencialidad en Documental.

5.6 Modelos de Datos (Referencia)

Se adoptan los modelos ER finales en Mermaid (.mmd) para SQL Server y MongoDB como referencia normativa para el diseño físico y la validación de integraciones. El inventario de entidades/tablas (5.1) y colecciones (5.2) se encuentra sincronizado con dichos modelos, incluyendo la auditoría transversal (audit_logs) y convenciones snake_case.

• SQL Server: dominios de contratos/áreas, GIS (catálogo de capas), sísmica (línea 2D, stack 2D, volumen 3D, procesamiento, modelos de velocidad) y pozos (pozo/ramal/registros).

• MongoDB: colecciones homólogas con validación JSON Schema, referencias a archivos (file_refs/location/checksum) y campos de auditoría.

• Reglas de evolución: versionado de esquema y RTM enlazado a cambios de modelo y APIs.

5.7 Stack Tecnológico y Despliegue Físico

Conforme al SAD v0.4, la plataforma BIEN adopta un modelo híbrido on-premise + nube privada con orquestación Kubernetes (K8s) para servicios de aplicación y pipelines, y SQL Server 2022 para datos estructurados con alta disponibilidad. Se integran servicios de observabilidad y seguridad de extremo a extremo, manteniendo soberanía del dato.

• Kubernetes: clúster con nodepools separados (sistema/aplicación), Ingress Controller (NGINX o AGIC), HPA para escalabilidad horizontal, despliegues blue/green y rollback automático.

CONFIDENCIAL

• Capa de datos estructurados: SQL Server 2022 con Always On Availability Groups (primario on-prem, réplica sincrónica y asíncrona para DR); índices columnstore, espaciales y full-text.

• Capa de datos no estructurados/semi-estructurados: MongoDB con validadores JSON y políticas de sharding/replicación; repositorio documental con checksum e integridad.

• Índice de búsqueda: motor dedicado (OpenSearch/Elasticsearch) para full-text, facetas y ranking; ingesta desacoplada por eventos.

• GIS: publicación OGC (WMS/WMTS/Vector tiles) y catálogo de productos/capas via map_product, map_layer_vector/raster, crs.

• Observabilidad: OpenTelemetry (trazas), Prometheus (métricas), Grafana (dashboards), SIEM/SOAR para correlación y respuesta.

• Seguridad: OIDC/SSO + MFA en frontera, mTLS interno, RBAC/ABAC por dominio/contexto, Vault/KMS para llaves y secretos (TDE/AES-256).

• Almacenamiento: NAS/SAN para respaldos y binarios pesados (raster, volúmenes sísmicos), con réplicas hacia sitio DR.

5.8 Integraciones Institucionales y Contratos API

Las integraciones se formalizan vía contratos REST/OGC para terceros y gRPC/HTTPS en el perímetro interno. Cada contrato incorpora versionamiento, límites de paginación, políticas de caché y trazabilidad (X-Correlation-ID).

• IdP/SSO: autenticación centralizada (OIDC/SAML) con validación de tokens en API Gateway/BFF.

• ESRI/Portal GIS: consumo y publicación de capas; sincronización de metadatos y SRID/CRS.

• BI/Reporting: exposición de datasets serving por JDBC/REST y entrega de reportes ejecutivos.

• Catálogo/Metadatos/Linaje: alta de activos, validación de esquemas y captura de linaje canónico.

• Auditoría: evento AccessLogged asíncrono al broker y persistencia WORM.

CONFIDENCIAL

6. COMPONENTES Y CAPACIDADES

6.1 Inventario de Tablas (SQL Server)

administrative_document, contract, contract_area, crs, dataset, dataset_file, document_set, map_layer_raster, map_layer_vector, map_product, seismic_2d_stack, seismic_3d_volume, seismic_line_2d, seismic_processing_project, seismic_survey, seismic_velocity_model, stratigraphic_unit, technical_report, unit_of_measure, well, well_log, well_log_curve_set, wellbore, core_sample, audit_logs.

6.2 Inventario de Colecciones (MongoDB)

administrative_document, audit_logs, commodity, contract, contract_area, core_sample, crs, dataset_file, document_set, map_layer_raster, map_layer_vector, map_product, seismic_2d_stack, seismic_3d_volume, seismic_line_2d, seismic_processing_project, seismic_survey, seismic_velocity_model, stratigraphic_unit, technical_report, unit_of_measure, well, wellbore, wellbore_marker_set, well_log, well_log_curve_set.

6.3 Zonas y Esquema de Almacenamiento

• Raw/Landing: ingreso directo, trazabilidad y validaciones mínimas.

• Curated: estandarización, enriquecimiento y conformado de entidades maestras.

• Serving: consumo (APIs, BI/Analítica, Búsqueda Unificada).

6.4 Ingesta y Procesamiento

• Patrones de integración batch/stream/evento (RF-07, RF-09) con orquestación de workflows.

• Validación de esquema (SQL/JSON Schema), control de calidad y checksums.

• Indexación para búsqueda unificada (RF-01) y vistas de servicio para analítica (RF-04).

CONFIDENCIAL

6.5 Exposición y Consumo

• APIs de dominio (contratos, GIS, sísmica, pozos, documental).

• Servicios de mapas (WMS/WMTS/Vector tiles) para GIS.

• Dashboards y reportes (RF-04) sobre datos conformados.

6.6 Parámetros de Servicio (SLI/SLO) y Capacidad

Se definen objetivos de servicio para garantizar experiencia y continuidad operativa. Los SLO se monitorean con telemetría nativa e informes periódicos de cumplimiento.

• Disponibilidad: ≥ 99.5% mensual para servicios críticos.

• DRP: RTO ≤ 60 minutos; RPO ≤ 15 minutos con promoción automática de réplicas y refresh de endpoints.

• Consulta: p95 < 3 s en búsquedas sobre índices (caché BFF TTL≈60 s, paginación y límites).

• Ingesta: validación de esquema y DQ en < 5 min por lote nominal; DLQ para rechazos.

• Capacidad: crecimiento proyectado (almacenamiento caliente y frío); tiering por temperatura; compresión y políticas ILM.

• Seguridad: rotación de llaves cada 90 días, mTLS interno, cifrado en reposo (TDE/AES-256).

6.7 Alineación con Modelos ER y Diccionario de Datos

Los modelos ER finales (SQL Server y MongoDB) y el diccionario de datos regulan el diseño físico. Se adopta convención snake_case, claves primarias *_id, y tabla/colección audit_logs para trazabilidad.

• SQL Server: tablas contract, contract_area, administrative_document, map_product, map_layer_vector/raster, seismic_line_2d, seismic_2d_stack, seismic_3d_volume, seismic_processing_project, seismic_velocity_model,

CONFIDENCIAL

well, wellbore, well_log, well_log_curve_set, core_sample, crs, stratigraphic_unit, unit_of_measure, technical_report, dataset, dataset_file, document_set y audit_logs.

• MongoDB: colecciones homólogas + wellbore_marker_set, commodity; validadores JSON (file_refs/location/checksum).

• Índices: por identificadores, fechas y campos de búsqueda; espaciales para geometrías y bbox en GIS.

• Auditoría: created_by/at, updated_by/at, status — transversales en modelos y bitácora central.

7. Operación y Gobierno

7.1 Gobierno de Datos

• Roles: Administrador de datos (gobernanza), Oficial de seguridad, Auditor SIEM, Operadores (GIS/Documental/BI) y Mesa de Ayuda.

• Catálogo/Linaje (RF-06) con reglas de calidad por dominio.

7.2 Seguridad y Cumplimiento

• Identidad y acceso por roles (RF-08).

• Auditoría consolidada (audit_logs) y controles de confidencialidad en documental.

• Backups cifrados y segregación de ambientes.

7.3 Respaldo y Recuperación

• Estrategias de backup para SQL y snapshots/export para Mongo/Documental.

• RPO/RTO por dominio y pruebas periódicas de restauración.

• Bitácora de ejecuciones y validaciones post-restore.

CONFIDENCIAL

7.4 Monitoreo y Observabilidad

• Métricas: disponibilidad, latencia de ingesta, éxito de jobs, uso de almacenamiento, salud de índices.

• Alertas proactivas y SIEM para eventos críticos.

7.5 Matriz RACI

La siguiente matriz RACI consolida responsabilidades por rol en las actividades críticas de infraestructura y datos. R = Responsible, A = Accountable, C = Consulted, I = Informed.

Actividad	Adm. de Datos (Gobernanza)	Oficial de Seguridad	Auditor SIEM	Operador GIS	Operador Documental	Operador BI/Analítica	Mesa de Ayuda	Arquitectura/SRE	Equipo de Desarrollo	PMO/PO	Áreas Técnicas ANH

CONFIDENCIAL

Definir arquitectura y estándares de datos	A	C	I	C	C	C	I	R	C	A	C
Implementar SQL Server & MongoDB	C	C	I	I	I	C	I	A	R	A	I
Gestión Documental & Metadatos	A	C	I	I	R	C	I	C	C	A	C
Seguridad IAM (SSO/MFA, RBAC/ABAC)	C	A	C	I	I	I	I	R	C	A	I
Auditoría y SIEM (logs/trazas)	I	C	A	I	I	I	I	R	C	A	I
Backups y DRP (RPO/RTO)	C	C	I	I	I	I	I	A	R	A	I
Monitoreo y Observabilidad	I	C	A	I	I	I	I	R	C	A	I
Pipelines de Ingesta & Integraciones	C	C	I	I	I	C	I	R	A	A	C
Exposición (APIs/Servicios/Mapas)	C	C	I	R	R	C	I	A	A	A	I
Cumplimiento contractual & evidencias	C	C	C	I	I	I	I	C	C	A	R
UAT, Go-Live controlado y estabilización	I	C	I	I	I	C	I	A	R	A	R

CONFIDENCIAL

7.6 Plan de Continuidad (Backups/Restore/DRP)

Se estandariza un plan de respaldo y recuperación por dominio, con pruebas de restauración y evidencia para auditoría.

• SQL Server: FULL/DIFF/LOG hacia NAS/SAN + replicación a almacenamiento externo; pruebas de restore trimestrales; verificación de TDE/llaves.

• MongoDB: snapshots y exports programados; restauración en entorno aislado para validación; verificación de validadores JSON.

• Documental/GIS: snapshots de repositorios y catálogo; rehidratación controlada y checksum.

• Runbooks: promotePrimary() para índice/BD; refreshEndpoints() en servicios; colas diferidas durante contingencias.

• Evidencias: bitácora de backups, resultados de restauración y verificación de RPO/RTO.

8. MAPA DE REQUERIMIENTOS – CAPACIDADES

• RF-01 Búsqueda Unificada - indexación multi-dominio, metadatos normalizados, vistas de búsqueda.

• RF-02 Gestión Documental - repositorio, metadatos, checksums, permisos y clasificación.

• RF-03 Geovisor GIS - map_product, map_layer_*, crs y servicios de mapas.

• RF-04 Analítica & Reportes - Serving y modelos conformados.

• RF-05 Asistente IA On-Prem - acceso controlado y auditoría.

• RF-06 Catálogo/Metadatos/Linaje - entidades de catálogo, legal_tags, acl, ancestry.

• RF-07 Integraciones/Workflow - orquestación y colas de trabajo.

• RF-08 Identidad/Acceso/Auditoría - IAM + audit_logs.

• RF-09 Ingesta Continua/Operación Evolutiva - pipelines y versionado de esquemas.

• RF-10 Organización Rondas/Licitaciones - entidades contractuales y documental.

CONFIDENCIAL

9. Consideraciones de Despliegue

• Ambientes: desarrollo, pruebas y producción; segregación de datos sensibles.

• Automatización: IaC para provisión y CI/CD para pipelines.

• Capacidad: crecimiento previsible por dominio (sísmica/pozos/GIS/documental).

9.1 RTM: Requerimientos - Capacidades/Secciones

La siguiente matriz resume la trazabilidad entre RF/NFR y las capacidades/artefactos de este documento.

• RF-01 (Búsqueda Unificada) - índice dedicado, 6.5, 6.6, 5.8.

• RF-02 (Gestión Documental) - repositorio documental, 5.2, 6.7.

• RF-03 (Geovisor GIS) - 5.3, crs y capas vector/raster, índices espaciales.

• RF-04 (Analítica & Reportes) - Serving/BI, 5.5, 5.8.

• RF-05 (Asistente IA On-Prem) - seguridad/observabilidad, catálogos y auditoría.

• RF-06 (Catálogo/Metadatos/Linaje) - 5.4, 6.7.

• RF-07 (Integraciones/Workflow) - 5.8, mensajería y DLQ.

• RF-08 (Identidad/Acceso/Auditoría) - 5.5, 6.5, 6.7.

• RF-09 (Ingesta Continua/Operación) - 6.4, 6.6.

• RF-10 (Rondas/Licitaciones) - entidades contractuales/documentales (5.1, 6.7).

• NFR (disponibilidad, DRP, seguridad, observabilidad, desempeño) - 6.6, 7.6, 5.7.

10. Anexos

• Inventarios de tablas y colecciones (Sección 5).

• Modelos ER Mermaid (.mmd) para SQL Server y MongoDB.

• Validadores JSON (colecciones) y definiciones SQL.

CONFIDENCIAL