

**POLÍTICA CORPORATIVA DE NUEVA ARQUITECTURA ANALÍTICA - DATALAKE, DELTALAKE, UNITY CATALOG, METADATA DRIVEN**

**PROGRAMA DE GOBIERNO DEL DATO**

22/08/2025

**Contenido**

[**1.** **INTRODUCCIÓN** 4](#_Toc206161024)

[**2.** **OBJETIVO** 4](#_Toc206161025)

[**3.** **ALCANCE** 4](#_Toc206161026)

[**4.** **GLOSARIO DE TÉRMINOS** 4](#_Toc206161027)

[**5.** **PRINCIPIOS FUNDAMENTALES** 5](#_Toc206161028)

[**6.** **COMPONENTES DE LA ARQUITECTURA ANALÍTICA** 5](#_Toc206161029)

[**7.** **LINEAMIENTOS DATALAKE** 6](#_Toc206161030)

[**7.1. ESQUEMA MEDALLÓN COPORATIVO** 6](#_Toc206161031)

[**7.1.1 Capa Bronce** 6](#_Toc206161032)

[**7.1.2 Capa Silver** 7](#_Toc206161033)

[**7.1.3 Capa Gold** 7](#_Toc206161034)

[**8.** **LINEAMIENTOS DE UNITY CATALOG** 7](#_Toc206161035)

[**8.1 Catálogo centralizado** 7](#_Toc206161036)

[**8.2 Gestión de accesos basado en roles** 7](#_Toc206161037)

[**8.3 Trazabilidad y linaje de datos** 8](#_Toc206161038)

[**8.4 Integración con procesos Metadata Driven** 8](#_Toc206161039)

[**8.5 Clasificación y sensibilidad de datos** 8](#_Toc206161040)

[**8.6 Roles y Permisos en Unity Catalog** 8](#_Toc206161041)

[**9.** **LINEAMIENTOS DE METADATA** 9](#_Toc206161042)

[**6.1 Metadatos Técnicos** 9](#_Toc206161043)

[**6.2 Metadatos de Negocio** 9](#_Toc206161044)

[**6.3 Metadatos Operacionales** 9](#_Toc206161045)

[**6.4 Gobierno Activo de Metadatos** 9](#_Toc206161046)

[**6.5 Uso de Metadatos para Procesos Metadata-Driven** 9](#_Toc206161047)

[**10.** **CICLO DE VIDA** 10](#_Toc206161048)

[**11.** **MONITOREO** 10](#_Toc206161049)

[**12.** **REFERENCIA** 10](#_Toc206161050)

[**13.** **CONTROL DE VERSIONES** 11](#_Toc206161051)

[**14.** **APROBACIÓN** 11](#_Toc206161052)

[**15.** **VIGENCIA** 11](#_Toc206161053)

1. **INTRODUCCIÓN**

La presente política establece los lineamientos para la implementación y uso de la arquitectura analítica corporativa basada en Data Lake, Delta Lake, Unity Catalog y un enfoque metadata-driven. Su propósito es asegurar que los datos corporativos sean gestionados bajo principios de calidad, interoperabilidad y gobierno, habilitando un entorno analítico confiable que soporte tanto el análisis operativo como la analítica avanzada y la inteligencia artificial.

1. **OBJETIVO**

Establecer los lineamientos corporativos para el diseño, gestión y control de la arquitectura analítica y del uso de metadatos en el Corporativo. Se busca garantizar que la información alojada en el Data Lake, Delta Lake y catálogos centralizados (Unity Catalog) sea confiable, segura, trazable y explotable bajo un modelo metadata-driven que potencie la analítica avanzada, inteligencia de negocio y ciencia de datos.

1. **ALCANCE**

Aplica a:

* Todas las fuentes de datos integradas al Data Lake corporativo (SAP, sistemas operativos, IoT, geológicos, ambientales, financieros, etc.).
* Plataformas analíticas soportadas en Databricks, Delta Lake, Unity Catalog y repositorios de metadatos.
* Todas las áreas de negocio, TI, analistas de datos, científicos de datos, ingenieros de datos, arquitectos y usuarios que consuman o gestionen información analítica en el Corporativo.

1. **GLOSARIO DE TÉRMINOS**

* **Unity Catalog:** Servicio de administración de datos en la arquitectura Lakehouse, que facilita el gobierno y la catalogación de datos.
* **ETL (Extract, Transform, Load):** Proceso de extracción, transformación y carga de datos desde múltiples fuentes hacia un sistema centralizado.
* **Metadatos:** Datos que describen otros datos, proporcionando contexto sobre su origen, formato y uso.
* **Data Catalog (Catálogo de Datos):** Herramienta que organiza y documenta el inventario de datos, facilitando su búsqueda y gestión.
* **Delta Lake:** Capa de almacenamiento optimizada para el procesamiento de datos en entornos Lakehouse.
* **Data Pipeline:** Flujo automatizado que mueve los datos desde una fuente hasta un destino, aplicando transformaciones si es necesario.
* **Data Lake:** Almacén de datos masivo diseñado para almacenar datos estructurados y no estructurados a gran escala.
* **Data Warehouse:** Sistema centralizado que almacena datos estructurados optimizados para consultas y análisis.
* **Esquema:** Estructura que define la organización y el formato de los datos dentro de una base de datos o sistema.
* **Particionamiento de Datos:** Técnica para dividir grandes volúmenes de datos en segmentos más manejables.
* **Databricks:** Plataforma de análisis de datos en la nube que permite implementar la arquitectura Lakehouse.
* **Arquitectura Medallón:** Una arquitectura de datos en capas utilizada principalmente en entornos de big data. Organiza los datos en tres capas principales: Bronce, Plata y Oro.

1. **PRINCIPIOS FUNDAMENTALES**

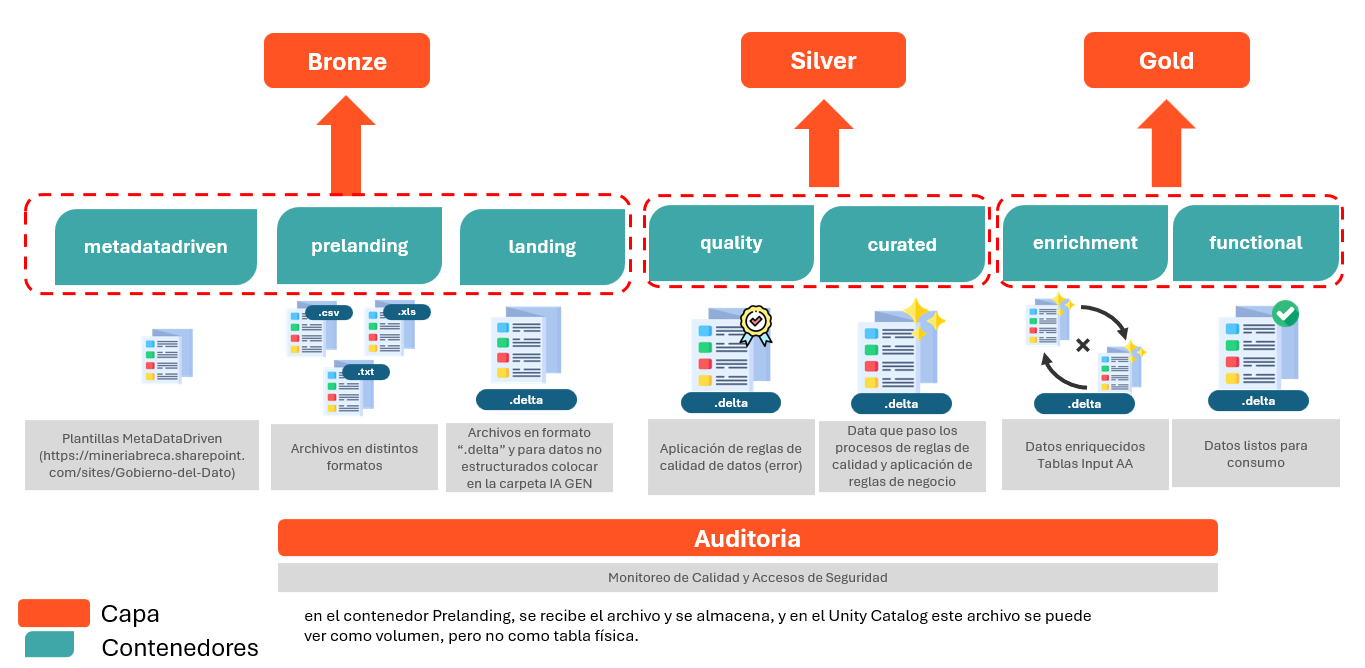
* **Arquitectura escalable y flexible:** Soportar crecimiento en volumen, variedad y velocidad de datos.
* **Metadata-driven:** Todo proceso analítico debe sustentarse en metadatos técnicos, de negocio y de calidad para reducir la dependencia manual.
* **Interoperabilidad:** Garantizar integración entre sistemas SAP, aplicaciones corporativas y plataformas cloud.
* **Trazabilidad y linaje:** Todos los datos deben tener registrado su origen, transformaciones y destino.
* **Seguridad y cumplimiento:** Accesos controlados por roles en Unity Catalog, con auditoría y segregación de privilegios.
* **Reutilización y estandarización:** Uso de patrones de arquitectura, modelos de datos y definiciones de metadatos comunes.

1. **COMPONENTES DE LA ARQUITECTURA ANALÍTICA**

* **Data Lake Corporativo:** Repositorio central de datos crudos y procesados.
* **Delta Lake:** Almacenamiento transaccional con control de versiones, calidad y consistencia de datos.
* **Unity Catalog:** Catálogo corporativo de datos y metadatos, con gestión de permisos, auditoría y clasificación de información.
* **Metadata Driven Framework:** Motor que habilita cargas, transformaciones, reglas de calidad y procesos analíticos basados en metadatos.

1. **LINEAMIENTOS DATALAKE**

## **7.1. ESQUEMA MEDALLÓN COPORATIVO**



La configuración de carpetas sugerida para el Corporativo modifica el esquema de medallón con una organización más exhaustiva para su utilización con Unity Catalog y Azure Data Lake. Las capas se reparten de la forma siguiente:

### **7.1.1 Capa Bronce**

El propósito de esta capa es almacenar los datos en su forma más cruda y original.  La información llega en diversos formatos y se almacena sin modificaciones.

* **Contenedor Metadata Driven:**

En este contenedor se da el uso de las plantillas, reglas, configuraciones y estructuras que se derivan o están parametrizadas por metadatos en el Meta Data Driven de MINSUR. Enlace de referencia: https://mineriabreca.sharepoint.com/sites/Gobierno-del-Dato

* **Contenedor Pre-Landing:**

Recibe archivos en distintos formatos estructurados (.parquet, .csv, .xlsx, .xml, json) sin procesar. Su propósito principal es preservar los datos en su forma más cruda para su reprocesamiento.

* **Contenedor Landing:**

Almacena archivos convertidos a formato delta y datos no estructurados (.txt, .pdf) en la carpeta "IA.GEN". En esta capa se realiza la unificación de formatos.

### **7.1.2 Capa Silver**

En esta capa los datos comienzan a ser procesados. Se aplican controles de calidad y reglas de negocio para garantizar que la información esté limpia y organizada.

* **Contenedor Quality:**

En este contenedor se aplican reglas de calidad de datos, se realizan mediciones para mejoras de calidad y se implementan procesos de remediación según criterios establecidos con las áreas de negocio.

* **Contenedor Curated:**

Aquí se integran y transforman los datos aplicando reglas de negocio para su estructuración adecuada. Los datos en esta capa están listos para análisis iniciales.

### **7.1.3 Capa Gold**

La información ya está depurada y enriquecida, lista para ser utilizada por diferentes aplicaciones y procesos analíticos. Es la capa donde los datos tienen mayor valor agregado, ya que están enriquecidos y estructurados.

* **Contenedor Enrichment:**

En esta capa los datos son enriquecidos mediante procesos adicionales, combinando información de múltiples fuentes e incrementando su valor analítico para que sean inputs de la Analítica Avanzada.

* **Contenedor Functional:**

Contiene datos completamente procesados y listos para el consumo (Tablas Delta).

### **7.1.4 Capa Auditoría**

La capa de Auditoría se entiende como un componente transversal que actúa sobre todas las capas (Bronze, Silver y Gold). No es un contenedor específico, sino un mecanismo de control y monitoreo continuo. La capa responsable de garantizar la trazabilidad, control y cumplimiento en el procesamiento de los datos a lo largo de todo el pipeline analítico. Se centra en dos dimensiones clave: calidad de datos y accesos y seguridad.

## **Organización de los datos en el Datalake**

* **Jerarquía:** la estructura de carpetas jerárquica separa las distintas capas del Lakehouse (ingesta, procesamiento, entrega).
* **Nomenclatura consistente:** convenciones de nomenclatura claras y coherentes para facilitar la identificación de datasets y su propósito.
* **Versionamiento de datos:** esquema de versionado para los datos, especialmente en la capa de procesamiento, para facilitar la auditoría y el seguimiento de cambios.
* **Registros de Metadatos:** un registro de metadatos adecuado para cada dataset, incluyendo descripciones, propietarios y etiquetas de gobernanza.

## **Lineamientos para Carga Full**

* Toda carga full debe registrarse en Unity Catalog como un nuevo commit de la tabla en Delta Lake, almacenando la versión y fecha de ejecución como metadatos de control.
* Los datos previos deben ser respaldados en la capa Bronze (histórico) antes de sobre escribir la versión activa en Silver o Gold.
* La frecuencia debe ser controlada por metadatos operacionales para evitar reprocesamientos innecesarias.
* Aplicar reglas de calidad en capa Silver para validar que la información cargada es íntegra y consistente.

## **Lineamientos para Carga Incremental**

* Debe implementarse siempre que la fuente permita identificar campos de control (fecha de modificación, hash, secuencia).
* Cada lote incremental debe ser procesado mediante un MERGE en Delta Lake, usando campos de control y claves de negocio para evitar duplicidades o sobreescrituras indebidas.
* Las cargas incrementales deben ser metadata-driven: el pipeline consulta metadatos para conocer el último punto de corte.
* Los datos rechazados en incrementales deben almacenarse en un área de remediación dentro de Silver.

1. **LINEAMIENTOS DEL DELTA LAKE**

Delta Lake será la capa transaccional y de gobernanza de datos sobre el Datalake Corporativo. Su adopción se establece bajo los siguientes lineamientos:

## **8.1 Estándar único de procesamiento**

* Todos los procesos de transformación, limpieza, enriquecimiento y consolidación de datos deberán implementarse sobre Delta Lake.
* Queda prohibido desarrollar procesos críticos de negocio fuera de esta capa.

## **8.2 Control de versiones y transacciones**

* Se habilitará el control de versiones (time travel) para asegurar la trazabilidad histórica y la recuperación de datos en caso de errores.
* Se garantizará atomicidad en operaciones de lectura y escritura, evitando inconsistencias.

## **8.3 Capas de datos estructuradas**

* Delta Lake organizará la información bajo el modelo Bronze, Silver y Gold:
* Cada capa tendrá controles de calidad, seguridad y auditoría específicos.

## **8.4 Calidad y gobernanza**

* Se implementarán reglas de validación automática en los procesos de escritura para asegurar integridad, unicidad y consistencia.
* Se documentará el linaje de los datos en integración con Unity Catalog.

## **8.5 Interoperabilidad**

* Delta Lake será el punto de integración entre SAP, aplicaciones satélites, sistemas externos y fuentes no estructuradas.
* Todo dataset integrado al ecosistema corporativo deberá pasar por Delta Lake para mantener la estandarización.

1. **LINEAMIENTOS DE UNITY CATALOG**

Unity Catalog será el componente central de gobierno de datos en la nueva arquitectura analítica de MINSUR. Su adopción se regirá bajo los siguientes lineamientos:

## **9.1 Catálogo centralizado**

* Será el repositorio único y corporativo de todos los objetos de datos (catálogos, esquemas, tablas, vistas, volúmenes, funciones, modelos).
* Ningún activo podrá ser creado o consumido fuera del control de Unity Catalog.

## **9.2 Gestión de accesos basado en roles**

* Se implementará un modelo de seguridad basado en el principio de mínimo privilegio.
* Los accesos se otorgarán de acuerdo con roles definidos corporativamente: Administrador, Data Engineer, Data Steward, Data Owner, Consumidor y Gobierno del Dato.
* La asignación de permisos será granular, diferenciando por ambiente (DEV, QA, PRD) y por tipo de objeto.

## **9.3 Nomenclatura Catálogo-Esquema-Tabla en Unity Catalog**

## **Catálogo:** *ambiente\_contenedor-capa\_empresa*

## Ejemplo: *dev\_bronze-landing\_minsur*

## **Esquema (Bronze - Silver):** *unidad minera\_fuente*

## Ejemplo: *Pisco\_Pisystem*

## **Esquema (Gold):** *unidad minera\_dominio*

## Ejemplo: *Pisco\_SuperintenedenciaTecnica*

## **Tabla:** *nombre de tabla*

## Ejemplo: *camas*

## **9.4 Trazabilidad y linaje de datos**

* Unity Catalog gestionará de manera central el linaje de datos, asegurando la trazabilidad desde la fuente hasta el consumo.
* Se registrarán metadatos técnicos y de negocio que permitan auditar el ciclo de vida de los datos.

## **9.5 Integración con procesos Metadata Driven**

* Unity Catalog se integrará con los procesos de ingestión y transformación de datos basados en metadatos.
* Permitirá estandarizar la nomenclatura, clasificar datos y habilitar la automatización de pipelines.

## **9.6 Clasificación y sensibilidad de datos**

* Se establecerán categorías de acuerdo a la clasificación de Seguridad de la Información (confidencial, interno, público) que condicionarán los niveles de acceso y controles aplicados.
* Unity Catalog será la única fuente oficial para consultar la clasificación y reglas de protección asociadas.

## **9.7 Roles y Permisos en Unity Catalog**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **N°** | **ROL** | **PERMISOS** | | | |
| **CATÁLOGO** | **ESQUEMA** | **TABLA** | **DESCRIPCIÓN** |
| 1 | Administrador Databricks | Use Catalog | Use Schema | Select  Insert  Update  Delete | Cuentan con privilegios para crear espacios de trabajo, configurar recursos en la nube, administrar usuarios, configuraciones y suscripciones. |
| Create Table |
| Modify Table |
| 2 | Administrador Workspace | Use Catalog | Use Schema | Select  Insert  Update | Pueden gestionar las identidades del Workspace, controlar el uso de recursos informáticos, y habilitar y delegar el control de acceso basado en roles. |
| Create Table |
| Modify Table |
| 3 | Administrador Metastore | Use Catalog | Use Schema | Select  Insert  Update | Cuentan con privilegios para administrar el almacenamiento de tablas y volúmenes de datos. |
| Create Table |
| Modify Table |
| 4 | Data Engineer | Use Catalog | Use Schema  Create Table  Modify Table | Select  Insert  Update  Delete | Pueden realizar cambios a nivel de Esquemas, permitiendo la creación y modificación de tablas de datos. |
| 5 | Data Science | Use Catalog | Use Schema | Select  Create View | Tienen permisos a nivel de tablas, permitiéndoles acceder a nivel de lectura a los datos y crear vistas personalizadas. |
| 6 | Business Data User | Use Catalog | Use Schema | Select | Los Business Data Users pueden acceder únicamente a las tablas de datos a nivel de lectura del dominio y subdominio asignado. |

1. **LINEAMIENTOS DE METADATA**

## **10.1 Metadatos Técnicos**

* Documentar estructura, tipo de dato, longitud, formato, reglas de validación y relaciones entre tablas.
* Registrar versión y cambios de esquemas en Delta Lake para asegurar la trazabilidad histórica.
* Mantener información de origen (sistema fuente, tabla, campo, fecha de extracción).

## **10.2 Metadatos de Negocio**

* Cada objeto de datos deberá contar con una definición de negocio clara y aprobada por el Data Steward del dominio.
* Se deben registrar reglas de calidad de negocio (campos obligatorios, dominios válidos, consistencia entre entidades).
* Todo dato debe estar asociado a un propietario de dominio (Domain Owner) y un responsable operativo (Data Owner).
* Clasificar los datos según criticidad, sensibilidad y relevancia estratégica.

## **10.3 Metadatos Operacionales**

* Incluir información de ejecución de pipelines: fecha, responsable, tiempo de carga, volumen procesado, errores.
* Documentar frecuencia de actualización (batch, streaming, tiempo real).
* Registrar indicadores de calidad aplicados y resultados obtenidos.

## **10.4 Gobierno Activo de Metadatos**

* Los metadatos deben mantenerse actualizados y auditados de manera continua.
* Los cambios en metadatos deberán pasar por el flujo de gestión de cambios de datos.
* La actualización de metadatos será responsabilidad compartida entre ingenieros de datos (parte técnica) y Data Stewards (parte de negocio).

## **10.5 Uso de Metadatos para Procesos Metadata-Driven**

* Toda carga, transformación o publicación de datos deberá configurarse a partir de metadatos, evitando parametrización manual.
* Los procesos de linaje, control de calidad y auditoría deberán habilitarse automáticamente mediante metadatos.
* El enfoque metadata-driven es obligatorio para la generación de reportes de trazabilidad y para habilitar modelos de autoservicio.

1. **CICLO DE VIDA**

* **Ingesta:** Todo dato ingresado al Data Lake debe ser descrito en Unity Catalog.
* **Transformación:** Se debe aplicar procesamiento controlado en Delta Lake, con reglas documentadas en metadatos.
* **Consumo:** Solo se podrán consumir datasets publicados con linaje completo y metadatos actualizados.
* **Retención y Depuración:** Se definirá la vigencia de datos en función de criterios regulatorios, de negocio y tecnológicos.

1. **MONITOREO**

* Métricas de calidad, uso y rendimiento de datos documentadas en el catálogo.
* Auditoría periódica de accesos y transformaciones.
* Alertas automáticas ante fallos en reglas de calidad o incumplimiento de políticas de metadatos.

1. **REFERENCIA**

* DG-POL-PCGD-MINS-001-Política Corporativa de Gobierno de Datos.
* DG-POL-PCCD-MINS-001-Política Corporativa de Calidad de Datos.
* DG-MAN-INGS-MINS-001-Procedimiento de Ingesta con Calidad y Seguridad de Datos.
* DG-PRO-PCCD-MINS-001-Procedimiento Corporativo de Calidad de Datos.

1. **CONTROL DE VERSIONES**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Versión** | **Fecha** | **Responsable** | **Cambios Realizados** | **Aprobado por** |
| 1.0 | 22/08/2025 | Jean Cabrera | Versión inicial | Comité Directivo |

1. **APROBACIÓN**

Esta política ha sido revisada y aprobada por el Comité Directivo y es de cumplimiento obligatorio para toda la organización.

[Nombre] Comité de Gobierno de Datos [Fecha]

1. **VIGENCIA**

Este lineamiento entra en vigor a partir de su aprobación formal y permanecerá vigente hasta su reemplazo o derogación.

**Aprobado por:** [Nombre del responsable]  
**Cargo:** [Cargo]  
**Fecha de aprobación:** [Fecha]  
**Versión:** 1.0  
**Fecha de próxima revisión:** [Fecha]

**APROBACION DEL DOCUMENTO**

|  |  |  |  |
| --- | --- | --- | --- |
| **Cargo** | **Nombre** | **Firma** | **Fecha** |
| **Supervisor De Sistemas De Información** | **Milagros Paico** |  |  |
| **Jefe De Arquitectura Y Desarrollo Ágil** | **Oscar Valencia** |  |  |
| **Jefe De Seguridad De La Información** | **Manuel Iturrizaga** |  |  |
| **Jefe de Analytics** | **Sthip Blas** |  |  |
| **Jefe Data Engineering** | **Luis Grados** |  |  |
| **Superintendente TI Marcobre** | **Marco Yacarini** |  |  |
| **Supervisor de Aplicación Marcobre** | **Jimmy Bendezú** |  |  |
| **Superintendente de TI & OT** | **Edwin Saldaña** |  |  |
| **Gerente de TI & OT** | **Dante Cárdenas** |  |  |
| **Gerente de TD** | **Piero Saravia** |  |  |