

Manual Maestro Data Lakehouse UNGRD

Equipo de Conocimiento UNGRD

2026-02-16

Table of contents

Preface	4
I PARTE I: FUNDAMENTOS Y ESTRATEGIA	5
1 Visión del Proyecto UNGRD	6
1.1 El Objetivo: Del Caos al Orden	6
1.2 ¿Qué es el Lakehouse?	6
1.3 Las Capas: La Analogía del Tratamiento de Agua	7
1.3.1 Capa Bronze (Agua Cruda)	7
1.3.2 Capa Silver (Agua Filtrada)	7
1.3.3 Capa Gold (Agua Potable/Premium)	7
1.4 Mapa de Herramientas (Ecosistema Alibaba Cloud)	7
1.5 Glosario Corporativo	8
II PARTE II: INGESTA Y ALMACENAMIENTO (OSS)	9
2 Gestión de Archivos en OSS	10
2.1 Flujo de Datos en el Lakehouse	10
2.2 Navegación por la Consola	11
2.3 Estándares de Naming: Regla UNGRD	12
2.4 Guía de Carga de Datos	12
2.4.1 Carga vía Consola Web (Paso a Paso)	12
2.4.2 Carga vía ossutil (Volúmenes masivos/Automatización)	13
2.5 Versionamiento y Recuperación	13
2.6 Ciclo de Vida: Optimización de Costos	13
III PARTE III: INTEGRACIÓN (DataWorks)	14
3 DataWorks - El Centro de Mando	15
3.1 Introducción a DataWorks	15

IV PARTE IV: PROCESAMIENTO (MaxCompute)	16
4 Procesamiento y SQL con MaxCompute	17
4.1 MaxCompute	17
V PARTE V: VISUALIZACIÓN	18
5 Quick BI	19
5.1 Visualización de Datos	19
6 DataV Geoespacial	20
6.1 Datos en el Mapa	20
VI PARTE VI: GOBERNANZA	21
7 Seguridad y Calidad de Datos	22
7.1 RAM y Gobernanza	22
7.2 DQC (Data Quality Control)	22

Preface

This is a Quarto book.

To learn more about Quarto books visit <https://quarto.org/docs/books>.

Part I

PARTE I: FUNDAMENTOS Y ESTRATEGIA

1 Visión del Proyecto UNGRD

1.1 El Objetivo: Del Caos al Orden

En la gestión del riesgo de desastres, el tiempo no es solo una métrica de rendimiento; es la diferencia entre la vida y la muerte. Históricamente, la **UNGRD** ha operado con datos fragmentados, silos de información y procesos manuales que retrasan la respuesta ante emergencias.

El proyecto **Data Lakehouse** nace con una misión clara: **Salvar vidas con datos**. Nuestra meta es transformar el caos de datos crudos en una fuente única de verdad, permitiendo que cada decisión —desde la entrega de ayuda humanitaria hasta la mitigación de riesgos de inundación— esté respaldada por evidencia analítica en tiempo real.

Nuestra Misión

Convertir la información en acción. Un dato bien procesado hoy es una vida salvada mañana.

1.2 ¿Qué es el Lakehouse?

El **Data Lakehouse** es la evolución de nuestra infraestructura tecnológica en Alibaba Cloud. Combina lo mejor de dos mundos:

1. **Data Lake**: La capacidad de almacenar volúmenes masivos de datos en su formato original (imágenes satelitales, sensores, reportes de campo) de manera económica y flexible usando **OSS**.
2. **Data Warehouse**: El rigor, la estructura y el rendimiento de una base de datos analítica de alto nivel usando **MaxCompute**.

Esta arquitectura híbrida nos permite ser ágiles sin perder la soberanía y gobernanza de los datos nacionales.

1.3 Las Capas: La Analogía del Tratamiento de Agua

Para entender cómo fluyen los datos en el Lakehouse, utilizamos la **Arquitectura Medallion**, que funciona exactamente como una planta de tratamiento de agua:

1.3.1 Capa Bronze (Agua Cruda)

Es el punto de entrada. Aquí almacenamos los datos tal cual llegan de las fuentes (como el agua tomada directamente de un río). Es completa pero contiene sedimentos e impurezas que no permiten su consumo directo. * **Estado:** Crudo / Sin procesar.

1.3.2 Capa Silver (Agua Filtrada)

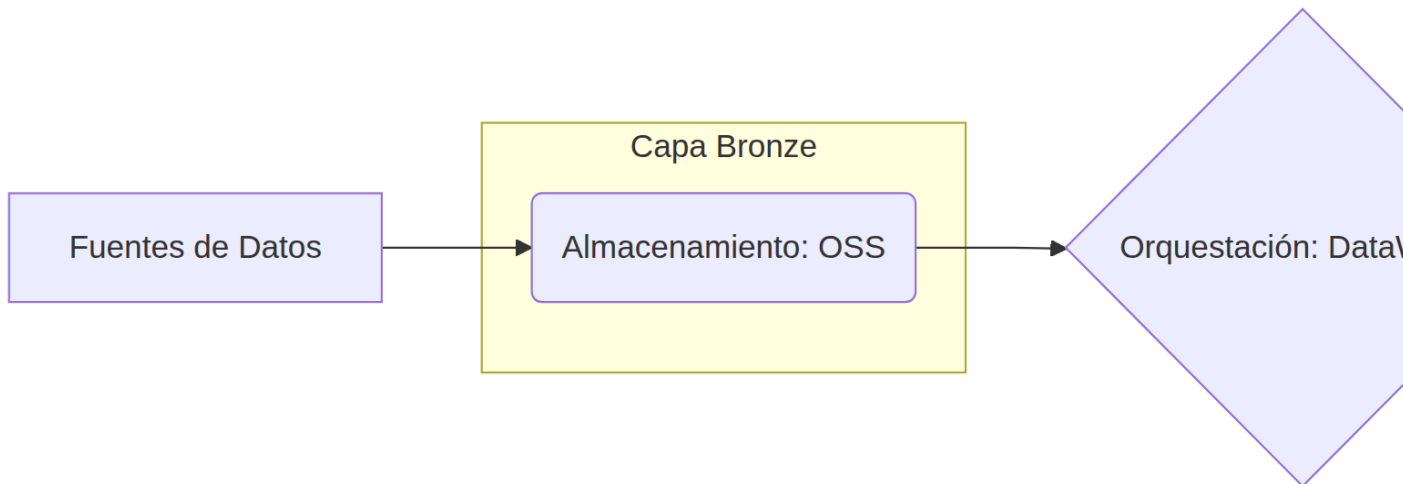
En esta etapa, limpiamos, filtramos y estandarizamos los datos. Eliminamos duplicados y corregimos errores. El dato ya es confiable para los analistas técnicos, similar a un agua que ha pasado por procesos de potabilización pero aún no está embotellada. * **Estado:** Limpio / Consistente.

1.3.3 Capa Gold (Agua Potable/Premium)

Es la capa final de valor. Aquí los datos están agregados y modelados para responder preguntas de negocio específicas (ej: “¿Cuántas familias fueron afectadas en Mocoa?”). Es el agua lista para el consumo masivo en dashboards de **Quick BI**. * **Estado:** Accionable / Optimizado para toma de decisiones.

1.4 Mapa de Herramientas (Ecosistema Alibaba Cloud)

La infraestructura se apoya en cuatro pilares fundamentales:



- **OSS (Object Storage Service):** Nuestro gran repositorio de archivos.
- **MaxCompute:** El motor que procesa millones de filas en segundos.
- **DataWorks:** El puente que conecta todas las herramientas y automatiza los flujos.
- **Quick BI / DataV:** Las ventanas a través de las cuales los directivos ven la realidad del territorio.

1.5 Glosario Corporativo

- **Arquitectura Medallion:** Metodología de organización de datos en capas (Bronze, Silver, Gold).
- **Lakehouse:** Plataforma que unifica capacidades analíticas y de almacenamiento masivo.
- **Ingesta:** El proceso de mover datos desde una fuente externa hacia nuestro sistema.
- **Gobernanza:** Reglas y controles para asegurar que el dato sea veraz, seguro y privado.
- **Ficha SUE:** Sistema Unificado de Información para la gestión del Riesgo.

Part II

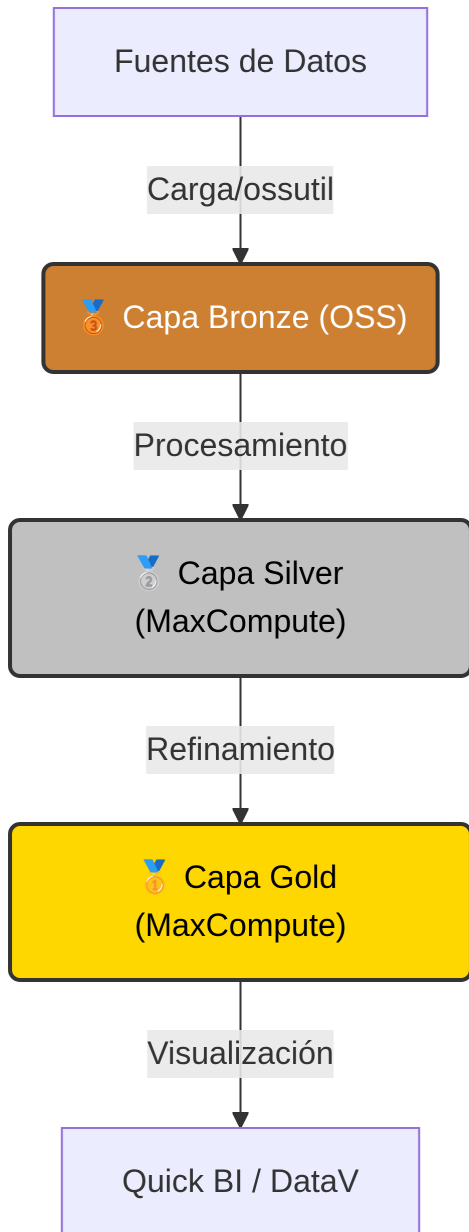
PARTE II: INGESTA Y ALMACENAMIENTO (OSS)

2 Gestión de Archivos en OSS

El **Object Storage Service (OSS)** de Alibaba Cloud es la base de nuestra capa **Bronze**. En este capítulo, aprenderás a gestionar los archivos de forma estandarizada y segura.

2.1 Flujo de Datos en el Lakehouse

Para contextualizar el rol de OSS, observemos cómo transitan los datos a través de las capas del proyecto:



2.2 Navegación por la Consola

Para gestionar los datos manualmente:

1. Inicia sesión en la consola de **Alibaba Cloud**.

2. En el menú de productos, busca y selecciona **Object Storage Service**.
3. En el panel izquierdo, haz clic en **Bucket List**.
4. Selecciona el bucket asignado a tu proyecto (ej: **ungrd-lakehouse-bronze**).
5. Usa la pestaña **Files** para navegar por las carpetas, subir o descargar archivos.

2.3 Estándares de Naming: Regla UNGRD

Para garantizar la trazabilidad y evitar colisiones de datos, **todos** los archivos cargados deben seguir esta estructura:

[PROYECTO]_[ZONA]_[FECHA]_[VERSION].[EXT]

Ejemplos: - SUE_MOCOA_20231025_V1.csv - SAT_VIGIA_20240112_V2.json

[!IMPORTANT] No uses espacios, tildes ni caracteres especiales en los nombres.
Usa siempre guion bajo (_) como separador.

2.4 Guía de Carga de Datos

2.4.1 Carga vía Consola Web (Paso a Paso)

Este método es ideal para archivos individuales o volúmenes pequeños de datos:

1. **Ingreso al Bucket:** Navega hasta tu bucket objetivo siguiendo los pasos de la sección 2.1.
2. **Preparación de Carpeta:** Entra en la carpeta de destino donde deseas que repose el dato (ej: **proyectos/sue/bronze/**).
3. **Iniciar Carga:** Haz clic en el botón azul **Upload** situado en la parte superior derecha.
4. **Selección de Archivos:**
 - Haz clic en **Select Files** para abrir el explorador de tu equipo.
 - O simplemente arrastra y suelta los archivos directamente en el recuadro gris.
5. **Configuración de Carga:**
 - **Upload To:** Verifica que la ruta de destino sea la correcta.
 - **File ACL:** Déjalo en **Inherited from Bucket** (esto asegura que herede los permisos de seguridad de la UNGRD).
 - **Storage Class:** Asegúrate de que esté en **Standard** para datos que van a ser procesados de inmediato.
6. **Finalizar:** Haz clic en el botón **Upload** al final del panel lateral. Podrás seguir el progreso en la barra de tareas inferior.

2.4.2 Carga vía ossutil (Volúmenes masivos/Automatización)

Para usuarios avanzados e ingenieros, recomendamos usar la herramienta de línea de comandos `ossutil`.

Comando básico de subida:

```
./ossutil64 cp ./mi_archivo_local.csv oss://ungrd-lakehouse-bronze/proyectos/sue/
```

2.5 Versionamiento y Recuperación

El Lakehouse tiene activado el **Bucket Versioning**. Esto significa que si sobrescribes un archivo con el mismo nombre, el sistema guarda una copia de la versión anterior.

[!WARNING] **Recuperación de Archivos Borrados:** Si borras un archivo accidentalmente, este no desaparece de inmediato. El sistema crea un “Delete Marker”. Para recuperarlo, debes activar la opción **Show Deleted Objects**, seleccionar la versión anterior al borrado y hacer clic en **Restore**.

2.6 Ciclo de Vida: Optimización de Costos

Los datos de desastres antiguos que ya no se consultan frecuentemente deben moverse a capas de almacenamiento más baratas.

1. **Standard:** Acceso frecuente, mayor costo.
2. **IA (Infrequent Access):** Acceso mensual, menor costo.
3. **Cold Archive:** Para datos históricos (de hace 5+ años). Los costos son mínimos, pero la recuperación puede tardar minutos u horas.

Configura reglas de **Lifecycle** en la pestaña **Basic Settings** para automatizar este movimiento y optimizar el presupuesto de la UNGRD.

Part III

PARTE III: INTEGRACIÓN (DataWorks)

3 DataWorks - El Centro de Mando

3.1 Introducción a DataWorks

Herramienta de orquestación y desarrollo de datos en el ecosistema Alibaba Cloud.

Part IV

PARTE IV: PROCESAMIENTO (MaxCompute)

4 Procesamiento y SQL con MaxCompute

4.1 MaxCompute

Motor de cómputo para análisis masivo de datos.

Part V

PARTE V: VISUALIZACIÓN

5 Quick BI

5.1 Visualización de Datos

Creación de dashboards interactivos para la toma de decisiones.

6 DataV Geoespacial

6.1 Datos en el Mapa

Visualización de eventos y riesgos en tiempo real mediante mapas interactivos.

Part VI

PARTE VI: GOVERNANZA

7 Seguridad y Calidad de Datos

7.1 RAM y Gobernanza

Gestión de identidades y recursos.

7.2 DQC (Data Quality Control)

Aseguramiento de la integridad de los datos en todas las capas.