



Practicum: Análisis y Gobierno de Datos ENSANUT 2018

Julio Cesar Velázquez Corona

Universidad Anáhuac Campus Norte

Enero - Mayo 2026



Asesora: Dra. María del Carmen Villar Patiño

Índice general

Capítulo 1 Preparación	1
1.1 Requisitos de Software	1
1.1.1 Instalación de Docker	1
1.1.2 Instalación de Ubuntu 20.04.6 LTS (Solo para Windows)	1
1.1.3 Instalación de OpenMetadata	2
1.2 Desarrollo Previo de Automatización de Creación y Llenado de Tablas	4
1.3 Gobernanza de Datos con OpenMetadata	5
1.3.1 Glosario	5
1.3.2 Clasificación	6
1.3.3 Métricas	8
1.4 OpenMetadata y otros Softwares	9
Glosario	10
Bibliografía	11

Capítulo 1 Preparación

1.1 Requisitos de Software

Un proyecto de ciencia de datos requiere diferentes tipos de software para llevarse a cabo y en el caso de nuestro proyecto, es necesario contar con los siguientes:

- Docker: es una plataforma para empaquetar y ejecutar aplicaciones en contenedores que llevan consigo las dependencias para que corran igual en distintas máquinas.
- Ubuntu 20.04.6 LTS: es un entorno Linux dentro de Windows que corre dentro de WSL (Windows Subsystem for Linux). Es básicamente el backend de Docker.
- OpenMetadata: es una plataforma unificada para el descubrimiento, la observabilidad y la gobernanza, impulsada por un repositorio central de metadatos, lineaje detallado y una colaboración fluida del equipo.

1.1.1 Instalación de Docker

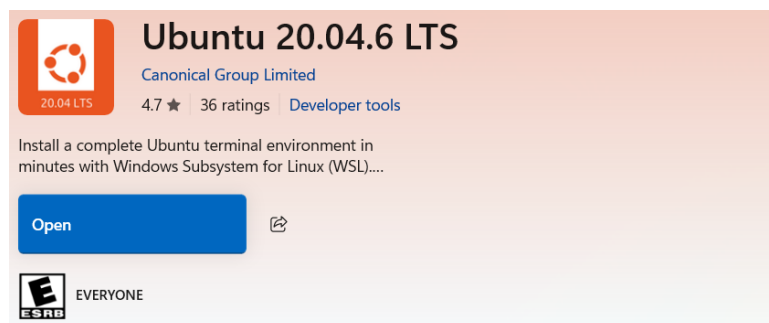
Accedemos a la [página de Docker](#) y damos click en **Download Docker Desktop**, luego aparecerá la opción de elegir el instalador que queremos dependiendo de nuestro SO. Ya con el instalador descargado, lo ejecutamos y seguimos los pasos que nos muestra. Cuando se indique, asegurarse de utilizar WSL 2 en lugar de Hyper-V. Por último, es necesario asignar como mínimo 6 GiB de memoria y 4 vCPUs. Para asegurarnos de que todo se instaló bien podemos correr en Powershell los siguientes comandos:

```
> docker --version
```

```
> docker compose version
```

1.1.2 Instalación de Ubuntu 20.04.6 LTS (Solo para Windows)

Puede descargarse directamente desde la Microsoft Store.



Ya instalado lo corremos y pedirá crear un usuario y una contraseña, luego entramos a Ubuntu e instalamos lo siguiente:

```
> sudo apt update
```

```
> sudo apt install python3-pip python3-venv
```

1.1.3 Instalación de OpenMetadata

En nuestro entorno Ubuntu en el caso de Windows o en la Docker CLI en el caso de Mac, colocar los siguientes comandos:

```
> mkdir openmetadata-docker && cd openmetadata-docker
> curl -sL -o docker-compose-postgres.yml https://github.com/open-metadata/
OpenMetadata/releases/download/1.11.3-release/docker-compose-postgres.yml
```

En el caso de Windows, entrar a Docker Desktop e ir a Settings > Resources > WSL integration y asegurarse de que todo esté activado de la siguiente manera:

Resources

Advanced File sharing Proxies Network WSL integration

Configure which WSL 2 distros you want to access Docker from.

☒ Enable integration with my default WSL distro

Enable integration with additional distros:

☒ Ubuntu-20.04

Refetch distros

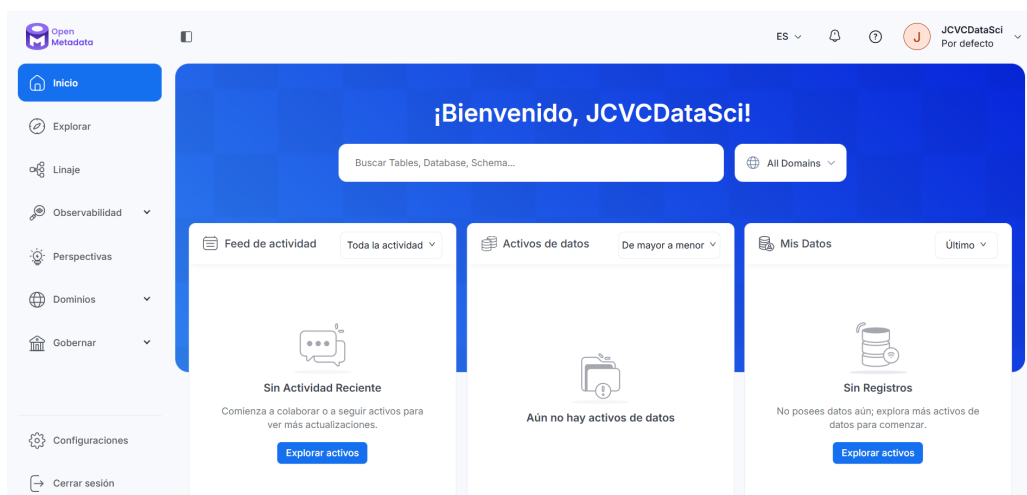
Colocamos el siguiente comando y al correrlo, reiniciamos la terminal:

```
> sudo usermod -aG docker $USER
```

Al volver a abrir la terminal corremos lo siguiente:

```
> docker compose -f docker-compose-postgres.yml up --detach
```

Al correr lo anterior se levanta el servicio de Docker Compose para OpenMetadata y podemos acceder al servicio desde <http://localhost:8585>. Ya dentro se nos pedirá crear una cuenta y ya podemos usar OpenMetadata.



Si posteriormente queremos volver a utilizar OpenMetadata hay que tener corriendo Docker Desktop en el fondo y correr los comandos que se presentarán a continuación:

```
> cd openmetadata-docker/  
> docker compose -f docker-compose-postgres.yml down  
> docker compose -f docker-compose-postgres.yml up -d
```

Finalmente acceder nuevamente a <http://localhost:8585>.








1.2 Desarrollo Previo de Automatización de Creación y Llenado de Tablas

El proceso que se debe llevar a cabo es el siguiente: **Automatizar la creación y llenado de las tablas, a una base de datos PostgreSQL y Oracle.** Se va a plantear cómo se tendría que realizar este proceso y de ahí se creará un diagrama de flujo para facilitar la reproducibilidad.






Nuestros datos se dividen en dos carpetas:

 nutricion	26/12/2025 11:28 a. m.	Carpeta de archivos
 salud	26/12/2025 11:29 a. m.	Carpeta de archivos

Cada una de esas carpetas tiene carpetas raíz con la siguiente estructura:

 conjunto_de_datos_cn_alimentos_adu_en...	26/12/2025 11:28 a. m.	Carpeta de archivos
 conjunto_de_datos_cn_alimentos_com_en...	26/12/2025 11:28 a. m.	Carpeta de archivos
 conjunto_de_datos_cn_alimentos_esc_ens...	26/12/2025 11:28 a. m.	Carpeta de archivos
 conjunto_de_datos_cn_alimentos_prees_e...	26/12/2025 11:28 a. m.	Carpeta de archivos
 conjunto_de_datos_cn_antropometria_en...	26/12/2025 11:28 a. m.	Carpeta de archivos
 conjunto_de_datos_cn_cat_alimentos_ens...	26/12/2025 11:28 a. m.	Carpeta de archivos
 conjunto_de_datos_cn_des_inf_ensanut_2...	26/12/2025 11:28 a. m.	Carpeta de archivos

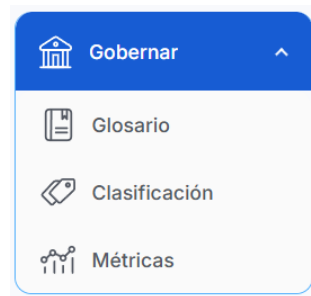
Cada carpeta raíz contiene lo siguiente:

 catalogos	26/12/2025 11:28 a. m.	Carpeta de archivos
 conjunto_de_datos	26/12/2025 11:28 a. m.	Carpeta de archivos
 diccionario_de_datos	26/12/2025 11:28 a. m.	Carpeta de archivos
 metadatos	26/12/2025 11:28 a. m.	Carpeta de archivos
 modelo_entidad_relacion	26/12/2025 11:28 a. m.	Carpeta de archivos

La carpeta **diccionario_de_datos** contiene un Excel con una descripción de cómo está estructurada la tabla que tiene los datos, la idea es que desde ahí se empieza la automatización de creación de tablas y por lo tanto nuestro primer paso es: leer los diccionarios de datos.

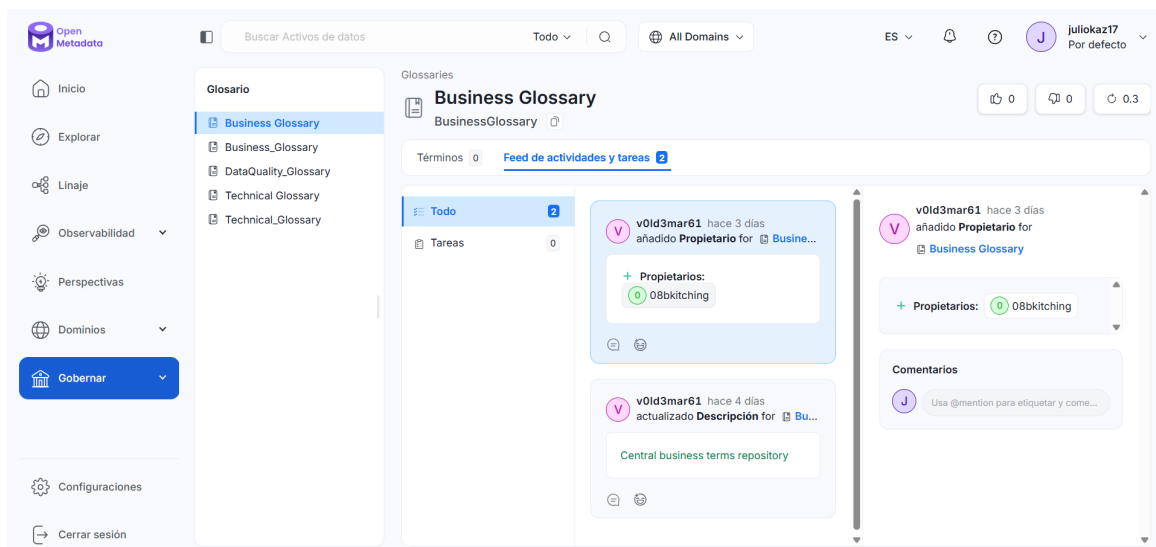
1.3 Gobernanza de Datos con OpenMetadata

El gobierno de datos es el conjunto de decisiones, reglas y prácticas con las que definimos y hacemos cumplir cómo se crean, se describen, se usan, se comparten, se protegen y se mantienen los datos dentro de una organización, para que sean confiables, seguros, consistentes y útiles para el negocio. En OpenMetadata tenemos tres artefactos que nos permiten llevar a cabo lo anterior y se encuentran en la sidebar del menú principal.



1.3.1 Glosario

En **Glosario** es donde se aloja el repositorio de los conceptos de los datos, eso quiere decir que ayuda a definir qué significa cada dato y también conectarlos con activos como tablas o columnas. En la sandbox nos encontramos con lo siguiente:



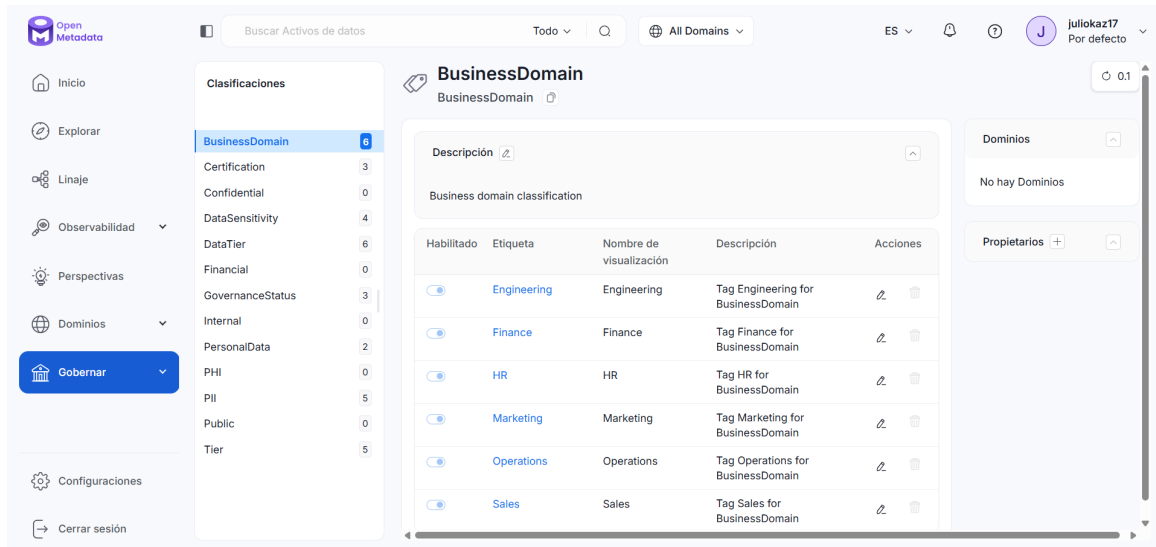
En el panel izquierdo podemos ver que hay diferentes glosarios y en este caso tenemos seleccionado el **Business Glossary**. A la derecha aparece nuestro título y vemos que en la parte de abajo dice **Términos 0**, eso quiere decir que aún no tenemos definiciones como tal de los datos. En la otra pestaña llamada **Feed de actividades y tareas** hay dos eventos registrados:

- El usuario **v0ld3mar61** añadió "Propietario" a nuestro glosario y el propietario asignado es **08bkitching**.
- Ese mismo usuario actualizó la descripción del glosario y la dejó como: *Central business terms repository*.

En la parte de la derecha aparece el cambio seleccionado con una sección de comentarios por si se quiere hablar del cambio.

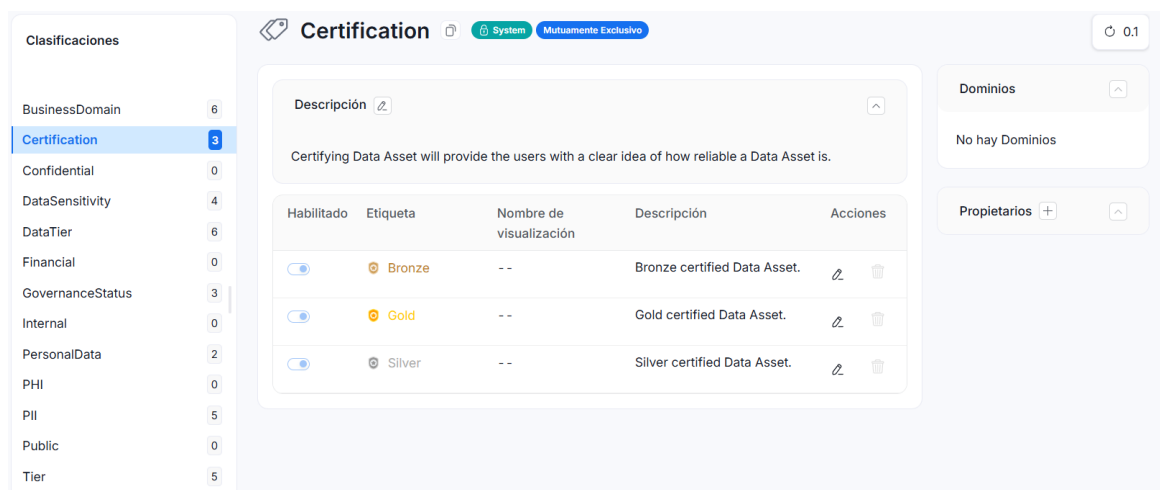
1.3.2 Clasificación

El siguiente artefacto es **Clasificación** y sirve para definir y administrar los datos que tenemos. No definimos el significado como tal, más bien se define qué tipo de dato es, cómo debe tratarse, si es PII o no PII, etc. La sandbox lo muestra de la siguiente forma:



En el sidebar debajo de **Clasificaciones** es donde se encuentra lo importante: son las etiquetas en las que pueden entrar nuestros datos. Algunas son creadas por default en el sistema y otras son creadas por usuarios. Para nuestro interes vale la pena describir las etiquetas creadas por el sistema:

- Certification: aquí se nos muestra si un dato es confiable o no.



- **PersonalData**: aquí se marcan datos que son personales.

PersonalData System Mutuamente Exclusivo 0.1

Descripción

Tags related classifying **Personal data** as defined by **GDPR**.

Note to Legal - This tag category is provided as a starting point. Please review and update the tags based on your company policy. Also, add a reference to your GDPR policy document in this description.

Habilitado	Etiqueta	Nombre de visualización	Descripción	Acciones
<input checked="" type="checkbox"/>	Personal	--	Data that can be used to directly or indirectly ... Ver más	🔗 🗑️
<input checked="" type="checkbox"/>	SpecialCategory	--	GDPR special category data is personal ... Ver más	🔗 🗑️

Dominios

No hay Dominios

Propietarios + [^](#)

- **PII**: aquí se etiqueta información personal que es capaz de identificar a una persona.

PII System Mutuamente Exclusivo 0.1

Descripción

Personally Identifiable Information that, when used alone or with other relevant data, can identify an individual.

Note to Legal - This tag category is provided as a starting point. Please review and update the tags based on your company policy. Also, add a reference to your PII policy document in this description.

Habilitado	Etiqueta	Nombre de visualización	Descripción	Acciones
<input checked="" type="checkbox"/>	Email	Email	Customer email address	🔗 🗑️
<input checked="" type="checkbox"/>	Name	Name	Customer name	🔗 🗑️
<input checked="" type="checkbox"/>	None	--	Non PII	🔗 🗑️

Dominios

No hay Dominios

Propietarios + [^](#)

- **Tier**: en esta etiqueta se clasifican nuestros activos de datos en capas que van del **Tier1** al **Tier5** y mientras más alta es la capa, más importante es el dato.

Tier System Mutuamente Exclusivo 0.1

Descripción

Tags related to tiering of the data. Tiers capture the business importance of data. When a data asset is tagged with Tier tag, all the upstream data assets used for producing it will also be labeled with the same tag. This will help upstream data asset owners to understand the critical purposes their data is being used.

Habilitado	Etiqueta	Nombre de visualización	Descripción	Acciones
<input checked="" type="checkbox"/>	Tier1	--	Critical Source of Truth business data assets of ... Ver más	🔗 🗑️
<input checked="" type="checkbox"/>	Tier2	--	Important business datasets for your ... Ver más	🔗 🗑️
<input checked="" type="checkbox"/>	Tier3	--	Department/group level datasets that are typical... Ver más	🔗 🗑️
<input checked="" type="checkbox"/>	Tier4	--	Team level datasets that are typically non-... Ver más	🔗 🗑️

Dominios

No hay Dominios

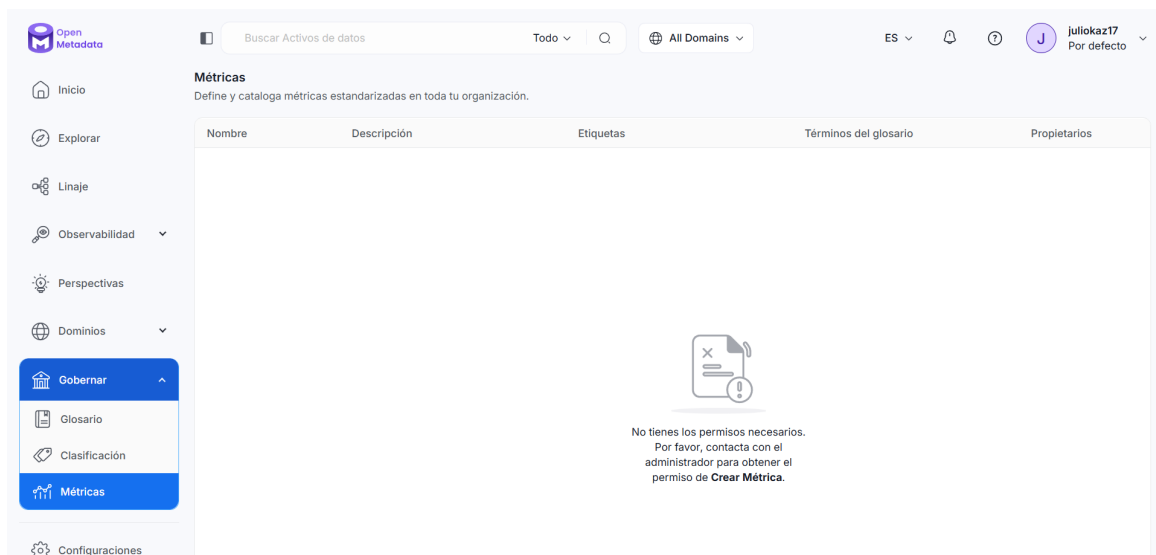
Propietarios + [^](#)

1.3.3 Métricas

Por último, en **métricas** es donde se define explícitamente con qué lógica opera una métrica definida. Por ejemplo: si tenemos un KPI llamado `monthly_arpu` (ingreso promedio por usuario al mes), podemos definir lo siguiente:

- Nombre o ID.
- Descripción.
- Tipo o unidad.
- Granularidad.
- Responsable.
- Glosarios vinculados.
- Tags.

En la sandbox desafortunadamente no podemos observar cómo es que se utiliza este artefacto porque no hay métricas definidas y aparte aparezco registrado como un usuario que no cuenta con los permisos necesarios para crear una.



1.4 OpenMetadata y otros Softwares

Hasta ahora, hemos descrito bastante OpenMetadata. Sin embargo, es de nuestro interés estar conscientes de que existe una competencia y que OpenMetadata no es la única opción. Los dos softwares más relevantes para lo que queremos y a mi parecer, los más "sonados" son Amazon Datazone y Oracle Cloud Infrastructure Data Catalog. A continuación se mostrará un cuadro comparativo de interés:

Software	OpenMetadata	Amazon Datazone	OCI Data Catalog
Instalación	Sí se instala y se mantiene con base de datos, plataforma y con un componente que hace las ingestas	No se instala porque ya viene como servicio pero sí hay que configurar acceso	No se instala pero hay que crear una instancia y configurar permisos
Experiencia recomendada	No mucha, solo en operación de plataformas y un poco en gobierno de datos	Más o menos, se requiere saber manejar el servicio IAM y un poco más de gobierno de datos	No mucha, solo estar envuelto en el entorno de Oracle
Ventajas	Más flexibilidad y control, podemos usar las herramientas que queramos	Está administrado por AWS y facilita la entrega de los datos	Está administrado por Oracle y brilla mucho en la cosecha (harvesting) de metadatos
Desventajas	Hay que mantenerlo constantemente y hay que hacer ajustes en el caso de actualizar algo	Disponibilidad solo en ciertas regiones y algunas cuotas en los servicios (pésimo)	Tiende a tener errores (bugs) según los known issues y también cuenta con algunas limitaciones

Glosario

AWS IAM Es un servicio que permite controlar de forma segura quién tiene acceso a qué recursos de AWS.

DDL Son las sentencias SQL para definir la estructura de la base de datos. O sea, crear, modificar y borrar objetos.

FKs Las llaves foráneas son restricciones que hacen que un campo en una tabla apunte a una fila existente en otra tabla, para mantener integridad referencial.

Harvesting Es el proceso de extraer, traer y actualizar metadatos desde las fuentes de datos hacia un catálogo o herramienta de gobierno.

Metadato Información que describe a un dato para que se pueda entender.

Bibliografía

- [1] C. V. Patiño. *Apuntes de Machine Learning*. Universidad Anáhuac México Campus Norte, 2025.