Actividad 1. Almacenes de datos y catálogo

**Introducción**

El Gobierno del Dato (Data Governance) es la gestión de la **disponibilidad**, **integridad**, **usabilidad** y **seguridad** de los datos utilizados en una empresa. Las grandes compañías consideran que un buen programa de Data Governance incluye un **consejo de gobierno**, un conjunto de **procedimientos** definidos, y un **plan** para ejecutar dichos procedimientos.

Un buen programa de gobierno de datos debe permitir a las partes interesadas comprender, proteger y confiar en los datos de una organización, especialmente a medida que las empresas crecen y acumulan más activos y **orígenes de datos**. Debido a esa acumulación exponencial de nuevos datos, las organizaciones se deben preocupar por definir entornos de Big Data adecuados para el almacenamiento y el acceso a los datos (datawarehouse, lagos de datos, entre otros). También debe preocuparse por diseñar una arquitectura de datos que ayude a gobernar esos orígenes de datos e integrarlos para que estén disponibles en toda la organización. Esta integración de datos es cada vez más importante, ya que afecta a los flujos de trabajo y la toma de decisiones de distintos equipos.

Si en una organización se define una buena estrategia para esa gobernanza, seguramente esta podrá contar con inventario detallado de qué datos tiene, dónde residen, cómo y quién los puede utilizar.

Hablar e intentar establecer una gobernanza de datos dentro de una organización es un reto muy ambicioso. Los programas de Data Governance pueden variar significativamente dependiendo de su enfoque (en cuanto a cumplimiento, integración de datos, Máster Data Management, etc.), es por ello por lo que, para esta actividad, nos centraremos concretamente en lo que respecta a las plataformas de datos (almacenamiento y el acceso a los datos). Una plataforma de gobierno de datos con un catálogo de datos integrado ayuda a su organización a encontrar, organizar, analizar, preparar y compartir datos para apoyar, por ejemplo, a sus iniciativas de IA. Las soluciones de gobierno de datos de las grandes plataforma (Databricks, Azure, IBM, etc.) justamente buscan garantizar que la interconexión de datos esté preparada para catalogar, proteger y gobernar los datos confidenciales y realizar un seguimiento de la filiación de datos.

**Objetivos**

Con esta actividad vas a aprender a trabajar con diferentes fuentes de datos y sobre ellas aplicar procesos que te permitan establecer varias estrategias para el acceso y la explotación de dichos datos.

Con el uso de herramientas especializadas, podrás crear espacios de trabajos para que diferentes usuarios puedan explotar los datos sin interferir con las tareas de otros usuarios y sin afectar la información en origen.

**Descripción**

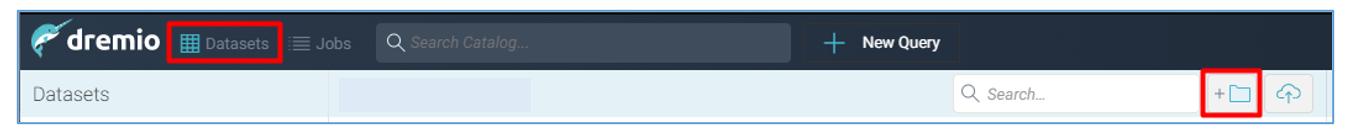
En el fichero comprimido *origenes\_de\_datos.zip* anexo a esta actividad, encontrarás cuatro (4) ficheros de datos con distintos formatos, extraídos de distintas sitios. Cada fichero representa un origen de dato diferente y, para esta actividad, tendrás inicialmente que organizarlos para su posterior análisis (la base de lo que, con muchos más orígenes de datos, podría ser un Data Lake).

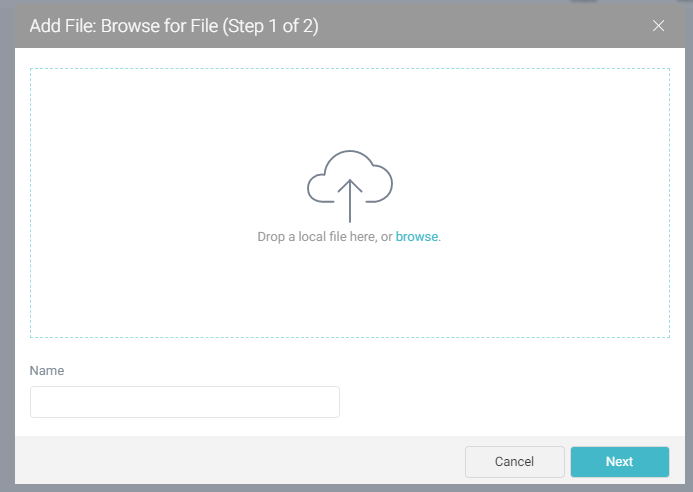
Una vez organizados, tendrás que crear espacios de trabajo donde construirás dataset para llevar a cabo tareas personalizadas para la explotación de dichos datos. Ten presente que los espacios de trabajo podrán estar asociados a diferentes usuarios, lo cual es un aspecto para destacar Dremio.

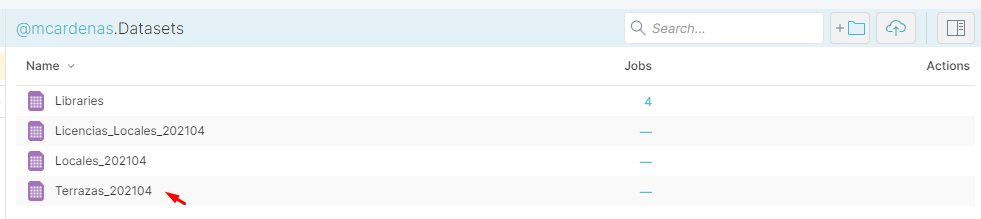
Los pasos que debes seguir para desarrollar esta actividad son los siguientes:

**Carga de datos**

1. Instala la herramienta propuesta para la actividad (consulta el anexo final).
2. Después de instalar la herramienta, utilízala para carga cada fichero. Crea una carpeta para almacenar todos los ficheros cargados (ahora serán los dataset).
   1. Comprueba que dicha carga se ha efectuado correctamente verificando que los datos están correctamente almacenados en los datasets.
   2. Al cargar cada fichero, realiza los ajustes correspondientes para que el fichero se almacene correctamente (encabezados, separadores, etc.).

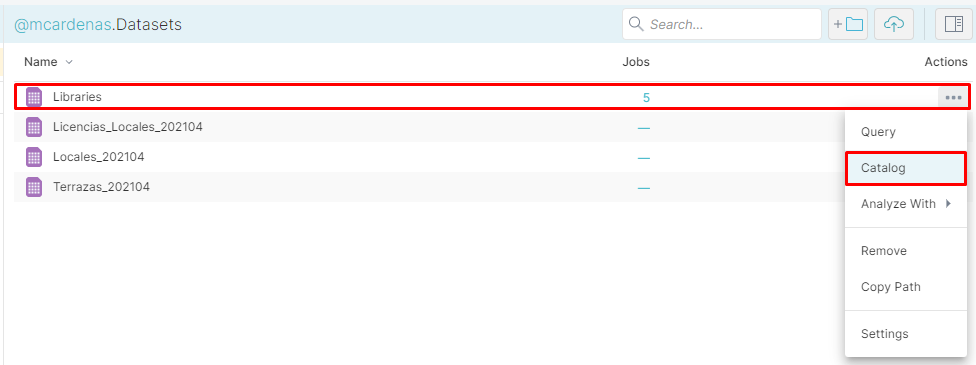


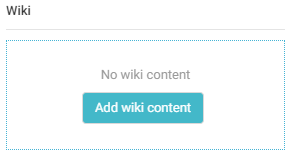


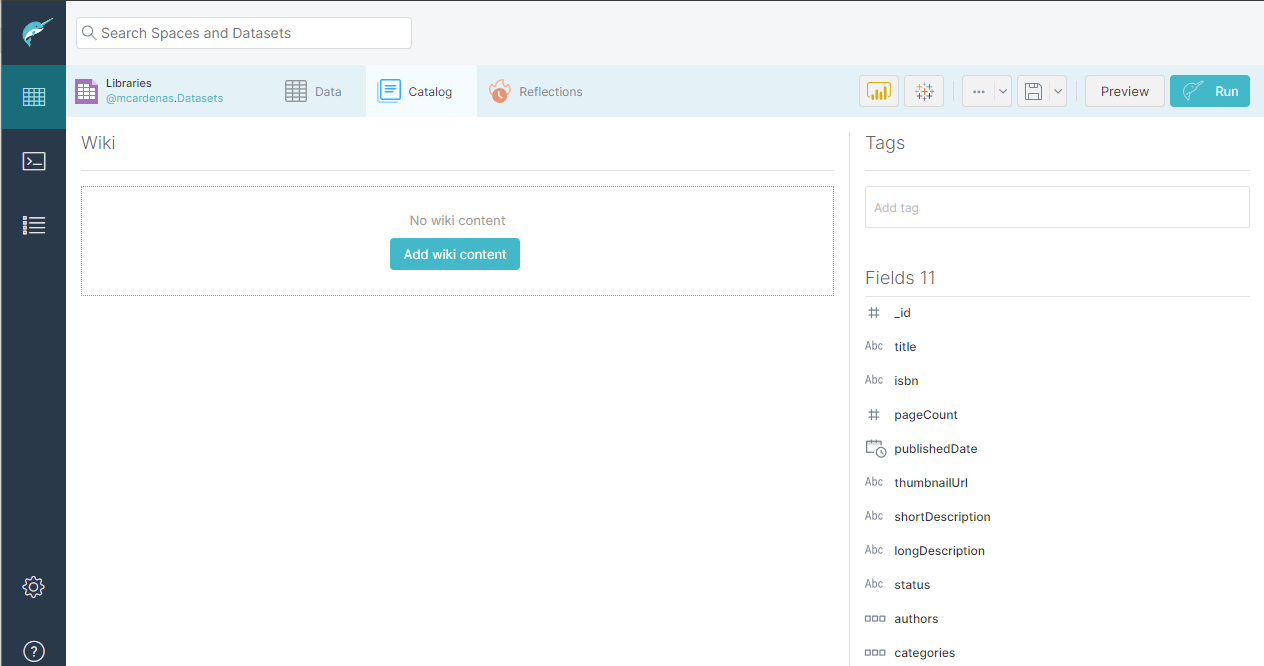


1. Por cada dataset tendrás que crear una “***wiki content***”. Esto consiste en una página que describe el dataset, la información que contiene y una lista de los campos que incluye (siéntete libre de incluir la información que consideres relevante).

En los casos que incorpores datos de una URL de Open Data (por ejemplo), puedes utilizar directamente la información que describe dicho fichero en el portal donde está alojado.

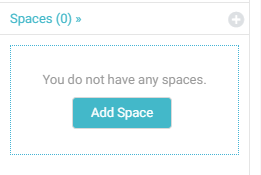


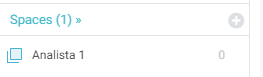


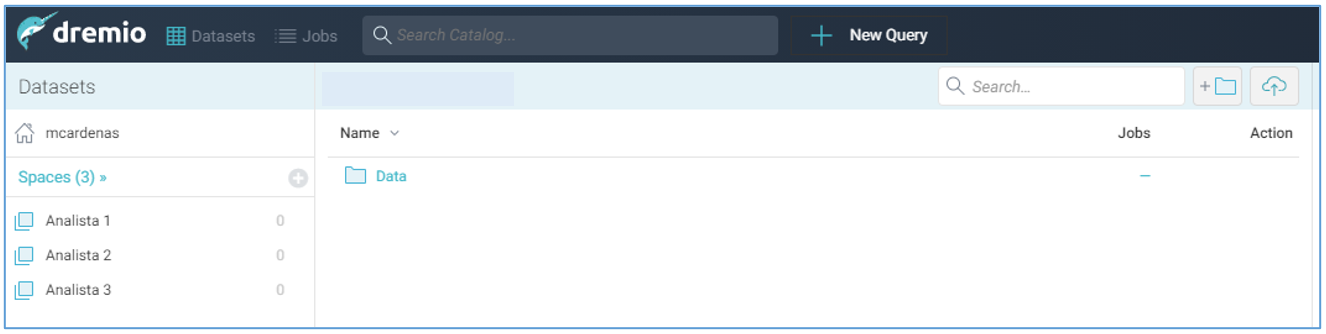


**Espacios de trabajo**

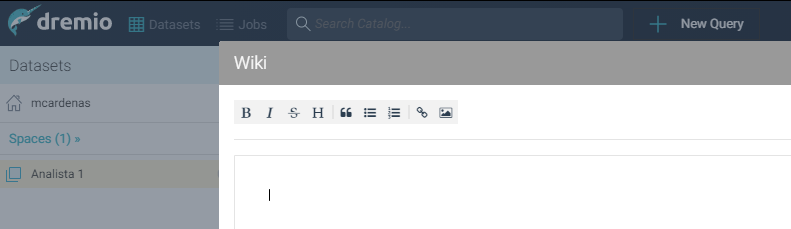
1. Sobre la herramienta debes crear 3 espacios de trabajos llamados:
   1. Analista 1
   2. Analista 2
   3. Analista 3





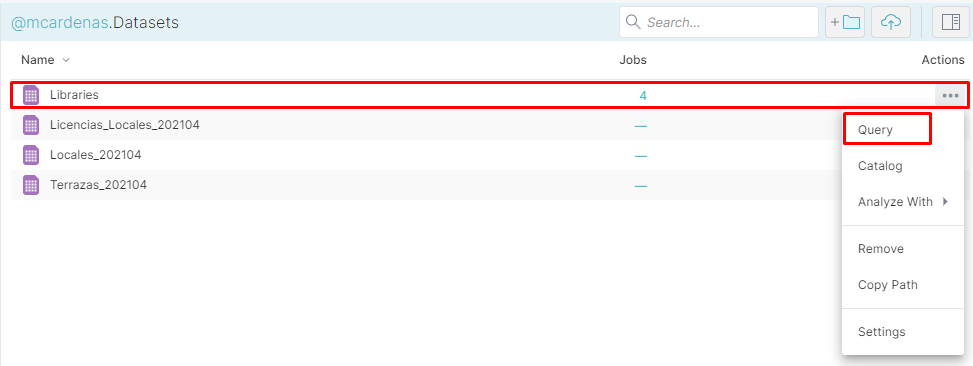


1. Cada espacio de trabajo debe llevar una “***wiki content***” que explique la finalidad del espacio de trabajo. Esto es una información libre que, también debe describir qué contiene el espacio de trabajo. Por ejemplo:
   1. **Analista 1**: agrupa datos relacionados con los ficheros CSV. Estos ficheros tienen que ver con la información de locales de la ciudad.
   2. **Analista 2**: agrupa datos relacionados con los ficheros JSON. Estos ficheros tienen información sobre las librerías de la ciudad.
   3. **Analista 3**: agrupa datos relacionados con los ficheros Open Data. Estos ficheros tienen información sobre el clima de la ciudad.

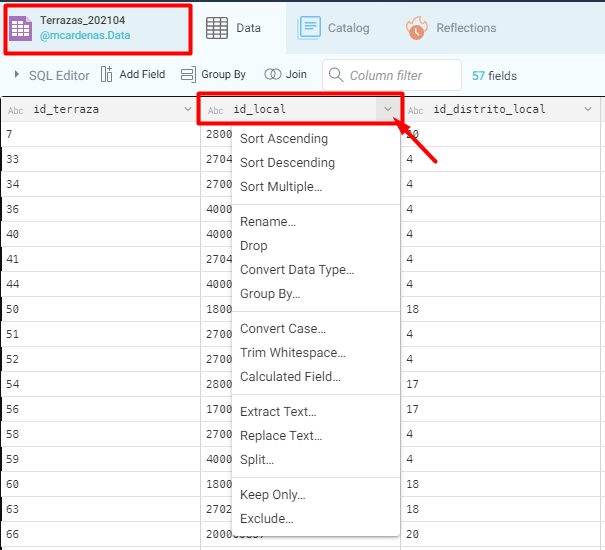


**Crear datasets personalizados**

Los dataset personalizados son **consultas** y **modificaciones** que aplicas sobre tu almacén de datos para luego publicarlos en los espacios de trabajo. Puedes crear tantas consultas como quieras y alojarlas en el espacio que consideres. Los analistas o científicos de datos trabajarán sobre los espacios de trabajo y no sobre los orígenes de datos como tal (los ficheros que has cargado). Esta es una de las principales cualidades de los Data Lake.



1. Abre el dataset **Terrazas\_202104** y realiza las siguientes modificaciones sobre él:
   1. Elimina todos los campos id\_\* excepto el campo id\_terraza.
   2. Elimina el campo “Escalera”.
   3. Crea un nuevo campo llamado Superficie\_TO que sume el campo Superficie\_ES y Superficie\_ES.
   4. Guarda la consulta con el nombre de **Terreza\_001** y guárdalo en el espacio **Analista 1.**
2. Abre el dataset **Licencias\_Locales\_202104** y elimina los campos del dataset excepto id\_local, ref\_licencia, desc\_tipo\_licencia, desc\_tipo\_situacion\_licencia y fecha\_dec\_lic. Guarda esta modificación con el nombre Licencias\_002 en el espacio de trabajo **Analista 1**.
3. Abre el dataset **Terrazas\_202104** y crea un join con el dataset Licencia\_002, utiliza el campo id\_local para hacer el inner join. Guarda esta modificación con el nombre Licencias\_Terrazas\_003 en el espacio de trabajo **Analista 2**.
4. Abre el dataset **books** y realiza los siguientes cambios:
   1. Elimina el campo \_id.
   2. Excluye los libros que no tienen ISBN (opción Exclude…, casilla *null*).
   3. En las columnas *authors* y *categories* aplica la opción unnest. ¿Comprendes qué ha ocurrido? Guarda esta modificación con el nombre Books\_ 001 en el espacio de trabajo **Analista 3**.



**Carga datos de ejemplos propios de Dremio (Opcional)**

Desde el repositorio de datos de ejemplo de Dremio, crea un espacio de trabajo con dichos datos, en concreto con el fichero *employees.parquet.*

1. Crea una Wiki Content para este nuevo repositorio.
2. Crea una consulta con el nombre *Empleados* y guárdala en un espacio de trabajo llamado “*Analista General*”.
3. Investiga qué son los ficheros *parquet* e indica una diferencia (la más mencionada) con respecto a los ficheros JSON.

**Rúbrica**

|  |  |  |  |
| --- | --- | --- | --- |
| Trabajo: Limpieza de dataset | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Criterio 1 | Apartado 1: instalación de la herramienta  Instalas la herramienta y la pones en ejecución. | 1 | 10% |
| Criterio 2 | Apartado 2: Carga de datos  - Desarrollas todos los puntos de este apartado. | 3 | 30% |
| Criterio 3 | Apartado 3: Espacio de trabajo  - Desarrollas todos los puntos de este apartado. | 2 | 20% |
| Criterio 4 | Apartado 4: dataset personalizados  - Desarrollas todos los puntos de este apartado. | 4 | 40% |
|  |  | **10** | **100 %** |

**Entrega**

Entrega un informe con los elementos habituales (portada, índice, nombre, fecha, numeración de páginas, etc.) en que des respuesta a los cuatro apartados planteados anteriormente. Por cada punto muestra una captura de pantalla que muestre la solución al problema planteado. Junto a cada captura realiza un breve comentario que explique lo que has hecho.

**Entrega en el Aula virtual**

Al finalizar la actividad, **harás la entrega del informe final** a través del apartado «Envío de actividades» del aula virtual. El informe debe ir con el nombre como se muestra a continuación:

* APELLIDO1\_APELLIDO2\_NOMBRE\_laboratorio\_1.pdf

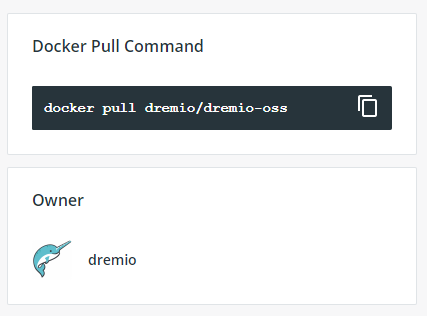
No uses tildes, ni apóstrofes ni ningún otro carácter que pudiera resultar conflictivo.

**Anexo: instalación**

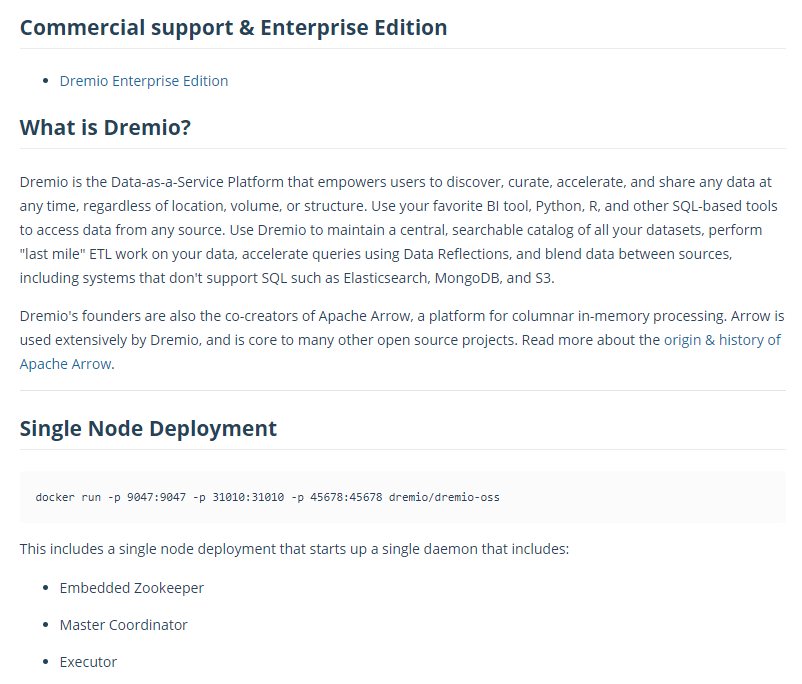
La herramienta propuesta para desarrollar la práctica es Dremio. Dremio es un motor de lago de datos de próxima generación que libera sus datos con consultas interactivas en vivo directamente en el almacenamiento del lago de datos en la nube. Para esta actividad lo usarás en modo On-Premise, es decir, tendrás que usarlo desde tu propio ordenador. La versión de instalación será a través de Docker, para ello se asume que tienes Docker instalado de asignaturas anteriores.

**Pasos de instalación**

1. Abre el CMD de Windows o Terminal de Linux.
2. Ejecuta la siguiente instrucción ([URL de consulta](https://hub.docker.com/r/dremio/dremio-oss/)):
   1. docker pull dremio/dremio-oss



1. Desde Docker Desktop, comprueba que se ha instalado la imagen de Dremio correctamente.
2. En la misma Terminal o CMD, ejecuta una instancia de Dremio con el siguiente comando:
   1. docker run -p 9047:9047 -p 31010:31010 -p 45678:45678 dremio/dremio-oss



1. Abre el navegador y accede a la URL <http://localhost:9047>
2. Crea una cuenta, asigna una contraseña y continúa con la práctica.

