Datos del estudiante

**Nombre y apellidos**

**Fecha de entrega**

Gobierno de Datos y Gestión de Datos con Herramientas de Microsoft

Objetivos de la actividad

Con esta actividad, desarrollarás habilidades para trabajar con diversas fuentes de datos y aplicar procesos que te permitirán diseñar múltiples estrategias para el acceso y aprovechamiento de estos datos.

Utilizando herramientas especializadas, aprenderás a crear entornos de trabajo donde diferentes usuarios puedan manipular y analizar los datos sin interferir en las tareas de otros ni comprometer la integridad de la información original.

Pautas de elaboración

Parte 1: Gobierno de Datos y Plataformas de Almacenamiento

Introducción

El Gobierno del Dato (*Data Governance*) se refiere a la administración de la disponibilidad, integridad, usabilidad y seguridad de los datos en una organización. Las grandes empresas consideran que un programa sólido de *Data Governance* debe incluir un comité de gobernanza, un conjunto de procedimientos claramente definidos y un plan de implementación para dichos procedimientos.

Un programa efectivo de gobierno de datos permite a las partes interesadas entender, proteger y confiar en los datos de la organización, especialmente a medida que las empresas crecen y acumulan más activos y fuentes de datos. Con el crecimiento exponencial de datos, las organizaciones deben preocuparse por establecer entornos de *Big Data* adecuados para el almacenamiento y acceso a los datos, como *data warehouses* y *data lakes*. Además, es crucial diseñar una arquitectura de datos que facilite la gobernanza de estas fuentes y su integración para que estén disponibles en toda la organización. Esta integración es vital ya que influye en los flujos de trabajo y en la toma de decisiones de diversos equipos.

Una estrategia bien definida de gobernanza de datos permitirá a una organización tener un inventario detallado de qué datos posee, dónde se encuentran, cómo se pueden utilizar y quién tiene acceso a ellos. Establecer una gobernanza de datos efectiva dentro de una organización es un reto significativo. Los programas de *Data Governance* pueden variar considerablemente dependiendo de su enfoque (cumplimiento, integración de datos, Gestión de Datos Maestros, etc.). En esta actividad, nos centraremos específicamente en las plataformas de datos relacionadas con el almacenamiento y acceso a los datos. Una plataforma de gobierno de datos con un catálogo de datos integrado ayuda a la organización a encontrar, organizar, analizar, preparar y compartir datos, apoyando iniciativas como la inteligencia artificial. Las soluciones de gobierno de datos de plataformas importantes (como Databricks, Azure, IBM, etc.) buscan garantizar que la interconexión de datos esté preparada para catalogar, proteger y gobernar datos confidenciales, además de rastrear su linaje.

En la segunda parte del laboratorio, os adentraréis en el emocionante mundo de la manipulación y gestión de datos utilizando Microsoft Visual Studio y SQL Server. La tarea consistirá en la transformación y carga de datos relacionados con la COVID-19, aprovechando estas poderosas herramientas, creando ETL. Se descargarán datos de ejemplo de COVID-19 de EE. UU. (si descargan otro tipo de datos, no hay problema), así como Microsoft Visual Studio y Microsoft SQL Server. A lo largo de este proceso, se seguirán las instrucciones proporcionadas en la documentación de estas herramientas, adquiriendo habilidades prácticas fundamentales en el ámbito de la informática y la gestión de datos.

* En el archivo comprimido origenes\_de\_datos.zip adjunto a esta actividad, encontrarás cuatro (4) archivos de datos en diferentes formatos, obtenidos de diversas fuentes. Cada archivo representa un origen de datos distinto. Para la primera parte de esta actividad, deberás organizarlos para su análisis posterior, sentando las bases de lo que, con más orígenes de datos, podría convertirse en un *Data Lake*.

Una vez organizados, crearás espacios de trabajo donde construirás conjuntos de datos para realizar tareas personalizadas de explotación de estos datos. Recuerda que los espacios de trabajo pueden estar asociados a diferentes usuarios, una característica destacada de Dremio.

* Además, deberán trabajar con procesos ETL para desarrollar habilidades en estas metodologías. Como se mencionó anteriormente, los datos seleccionados no son lo más relevante (se sugieren los datos mencionados anteriormente), sino el trabajo con las herramientas aquí descritas.

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

**Objetivo 1**

**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).

* Comprueba que dicha carga se ha efectuado correctamente verificando que los datos están correctamente almacenados en los *datasets*.
* Al cargar cada fichero, realiza los ajustes correspondientes para que el fichero se almacene correctamente (encabezados, separadores, etc.).

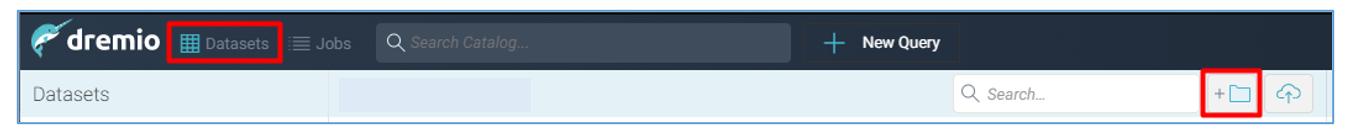


Figura 1. Fuente: elaboración propia.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Figura 2. Fuente: elaboración propia.

Tabla

Descripción generada automáticamente con confianza media

Figura 3. Fuente: elaboración propia.

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.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Figura 4. Fuente: elaboración propia.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Figura 5. Fuente: elaboración propia.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Figura 6. Fuente: elaboración propia.

**Espacios de trabajo**

1. Sobre la herramienta debes crear tres espacios de trabajos llamados:

* Analista 1.
* Analista 2.
* Analista 3.

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

Figura 7. Fuente: elaboración propia.

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza baja

Figura 8. Fuente: elaboración propia.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Figura 9. Fuente: elaboración propia.

2. Cada espacio de trabajo debe llevar una ***wiki content*** que explique la finalidad del espacio de trabajo (en el caso en que su versión de Dremio no lo permita, incluya el Wiki Content en el documento de la actividad). Esto es una información libre que, también debe describir qué contiene el espacio de trabajo. Por ejemplo:

* **Analista 1**: agrupa datos relacionados con los ficheros CSV. Estos ficheros tienen que ver con la información de locales de la ciudad.
* **Analista 2**: agrupa datos relacionados con los ficheros JSON. Estos ficheros tienen información sobre las librerías de la ciudad.
* **Analista 3**: agrupa datos relacionados con los ficheros Open Data. Estos ficheros tienen información sobre el clima de la ciudad.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Figura 10. Fuente: elaboración propia.

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*.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Figura 11. Fuente: elaboración propia.

1. Abre el dataset **Terrazas\_202104** y realiza las siguientes modificaciones sobre él:

* Elimina todos los campos id\_\* excepto el campo id\_terraza.
* Elimina el campo Escalera.
* Crea un nuevo campo llamado Superficie\_TO que sume el campo Superficie\_ES y Superficie\_ES.
* 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:

* Elimina el campo \_id.
* Excluye los libros que no tienen ISBN (opción Exclude…, casilla *null*).
* 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**.

Interfaz de usuario gráfica, Tabla

Descripción generada automáticamente

Figura 12. Fuente: elaboración propia.

Nota a tener en cuenta

Si durante la carga de alguno de los archivos existiese alguna inconsistencia de datos, se deberán corregir utilizando lenguaje de programación Python garantizando que todas las líneas del fichero estén introducidas de forma correcta y uniforme.

Por ejemplo, si una de las columnas es un de tipo integer, pero a partir de un punto hay inconsistencia, se utilizará Python para corregir esto y posteriormente realizar la carga en Dremio de manera adecuada.

Se añadirá al documento de la actividad un Jupyter Notebook donde se realice la corrección.

**Objetivo 2**

**Tarea de realización de ETL y modelado de datos**

En esta segunda parte de la actividad, tendrás que llevar a cabo la elaboración de ETL usando Microsoft Visual Studio. Posteriormente, volcarás estos datos en SQL Server y una vez ahí, se construirá el modelo de datos que corresponda, según lo visto en clase.

Para ello, tendrás que descargar los datos de COVID-19 disponibles en fuente abierta de Microsoft (no te preocupes tanto de los datos, si no de utilizar las herramientas). También tendrás que descargar Microsoft Visual Studio y SQL Server. Seguirás las indicaciones mostradas en las respectivas documentaciones.

Pasos:

* 1. Descarga de datos Ejemplo de COVID-19 de EE. UU.
* 2. Descargar Microsoft Visual Studio.
* 3. Descargar Microsoft SQL Server.
* 4. Realización de ETL desde Excel a SQL Server y construcción del modelo de datos.

**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.*

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

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/)):

* docker pull dremio/dremio-oss

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Figura 13. Fuente: elaboración propia.

3. Desde Docker Desktop, comprueba que se ha instalado la imagen de Dremio correctamente.

4. En la misma Terminal o CMD, ejecuta una instancia de Dremio con el siguiente comando:

* docker run -p 9047:9047 -p 31010:31010 -p 45678:45678 dremio/dremio-oss

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Figura 14. Fuente: elaboración propia.

5. Abre el navegador y accede a la URL <http://localhost:9047>

6. Crea una cuenta, asigna una contraseña y continúa con la práctica.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente con confianza media

Figura 15. Fuente: elaboración propia.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Figura 16. Fuente: elaboración propia.

Extensión y formato

Entrega un informe con los elementos habituales (portada, índice, nombre, fecha, numeración de páginas, etc.) en que des respuesta a los apartados planteados anteriormente, todo explicado, si no, penalizará en la puntuación. 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. Recuerda la entrega del JupyterNotebook en caso de que fuese necesario explicando cómo se ha solventado el problema.