****

**Universidad Internacional de La Rioja**

**Facultad de Ingeniería y Tecnología**

# **Máster Universitario en Análisis y Visualización de Datos Masivos / Visual Analitics & Big Data**

**Laboratorio 1**

|  |  |
| --- | --- |
| **Actividad de estudio presentado por:** | **Juan David Escobar Escobar** |
| **Tipo de trabajo:** | **Actividad** |
| **Modalidad:** | **Individual** |
| **Director/a:** | **Xiomara Blanco** |
| **Fecha:** | **Mayo 2022** |

**Índice de contenidos**

Table of Contents

[Máster Universitario en Análisis y Visualización de Datos Masivos / Visual Analitics & Big Data 1](#_Toc100682399)

[1. Carga de datos 3](#_Toc100682400)

[2. Espacios de trabajo 9](#_Toc100682401)

[2.1. Sobre la herramienta debes crear 3 espacios de trabajos llamados: 9](#_Toc100682402)

[ Analista 1 9](#_Toc100682403)

[ Analista 2 9](#_Toc100682404)

[ Analista 3 9](#_Toc100682405)

[2.2. 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: 9](#_Toc100682406)

[3. Crear datasets personalizados 10](#_Toc100682407)

[3.1. 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. 10](#_Toc100682408)

[3.2. Abre el dataset Terrazas\_202104 y realiza las siguientes modificaciones sobre él: 10](#_Toc100682409)

[a. Elimina todos los campos id\_\* excepto el campo id\_terraza. 11](#_Toc100682410)

[b. Elimina todos los campos Escalera 11](#_Toc100682411)

[c. Crea un nuevo campo llamado Superficie\_TO que sume el campo Superficie\_ES y Superficie\_ES. 11](#_Toc100682412)

[d. Guarda la consulta con el nombre de Terreza\_001 y guárdalo en el espacio Analista 1. 12](#_Toc100682413)

[3.3. 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. 12](#_Toc100682414)

[3.4. 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. 13](#_Toc100682415)

[3.5. Abre el dataset books y realiza los siguientes cambios: 13](#_Toc100682416)

[4. Carga de datos de ejemplo propias de Dremio (Opcional) 15](#_Toc100682417)

# **Carga de datos**

* 1. Instala la herramienta propuesta para la actividad (consulta el anexo final).

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

* 1. Después de instalar la herramienta, utilízala para carga cada fichero. Crea una carpeta para almacenar todos los ficheros cargados (ahora serán los dataset).
  2. Comprueba que dicha carga se ha efectuado correctamente verificando que los datos están correctamente almacenados en los datasets.

Los Dataset se suben sin problemas, excepto el Dataset libraries.json, el cual presenta un error de sintaxis en el campo\_id:

Graphical user interface, text, application

Description automatically generated

Se corrige el error identificado a través de un editor de consultas de MongoDB y por ultimo y se logran insertar los datos en MongoDB. Posteriormente al momento de subir el archivo a la plataforma Dremio surge un nuevo inconveniente con respecto al tipo de dato del campo \_id, el cual identifica el tipo de dato int64 como factor común, algunos valores están como tipo diccionario, objeto o Json, por lo cual se opta por remplazar estos valores por el consecutivo numérico que se identifica en la secuencia.

Text

Description automatically generated

Graphical user interface, application

Description automatically generated

* 1. Al cargar cada fichero, realiza los ajustes correspondientes para que el fichero se almacene correctamente (encabezados, separadores, etc.).

Graphical user interface, text, application

Description automatically generated

* 1. Parar 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.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

# **Espacios de trabajo**

# Sobre la herramienta debes crear 3 espacios de trabajos llamados:

# Analista 1

# Analista 2

# Analista 3

Graphical user interface, application

Description automatically generated

# 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:

* **Analista 1**: agrupa datos relacionados con los ficheros CSV. Estos ficheros tienen que ver con la información de locales de la ciudad.

Graphical user interface, application, Teams

Description automatically generated

* **Analista 2**: agrupa datos relacionados con los ficheros JSON. Estos ficheros tienen información sobre las librerías de la ciudad.

Graphical user interface, application

Description automatically generated

* **Analista 3**: agrupa datos relacionados con los ficheros Open Data. Estos ficheros tienen información sobre el clima de la ciudad.

Graphical user interface, text, application

Description automatically generated

# **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.

Graphical user interface, application

Description automatically generated

# Abre el dataset Terrazas\_202104 y realiza las siguientes modificaciones sobre él:

# Elimina todos los campos id\_\* excepto el campo id\_terraza.

**SELECT** id\_terraza, desc\_distrito\_local, desc\_barrio\_local,  clase\_vial\_edificio, desc\_vial\_edificio, nom\_edificio, num\_edificio, Cod\_Postal, coordenada\_x\_local, coordenada\_y\_local,desc\_tipo\_acceso\_local, desc\_situacion\_local, secuencial\_local\_PC, Escalera, coordenada\_x\_agrupacion, coordenada\_y\_agrupacion, rotulo, desc\_periodo\_terraza,  desc\_situacion\_terraza, Superficie\_ES, Superficie\_RA, Fecha\_confir\_ult\_decreto\_resol,DESC\_CLASE, DESC\_NOMBRE, nom\_terraza, num\_terraza, cal\_terraza, desc\_ubicacion\_terraza, hora\_ini\_LJ\_es, hora\_fin\_LJ\_es, hora\_ini\_LJ\_ra, hora\_fin\_LJ\_ra, hora\_ini\_VS\_es, hora\_fin\_VS\_es, hora\_ini\_VS\_ra, hora\_fin\_VS\_ra, mesas\_aux\_es, mesas\_aux\_ra, mesas\_es, mesas\_ra, sillas\_es, sillas\_ra

**FROM** "@juadaves".Datasets.Terrazas\_202104

Graphical user interface, text, application

Description automatically generated

# Elimina todos los campos Escalera

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

# Crea un nuevo campo llamado Superficie\_TO que sume el campo Superficie\_ES y Superficie\_ES.

**SELECT** id\_terraza, CAST(REPLACE("Superficie\_ES", ',', '.') **AS** **FLOAT**) + CAST(REPLACE("Superficie\_ES", ',', '.') **AS** **FLOAT**) **AS** "Superficie\_TO ", desc\_distrito\_local, desc\_barrio\_local, clase\_vial\_edificio, desc\_vial\_edificio, nom\_edificio, num\_edificio, Cod\_Postal, coordenada\_x\_local, coordenada\_y\_local, desc\_tipo\_acceso\_local, desc\_situacion\_local, secuencial\_local\_PC, coordenada\_x\_agrupacion, coordenada\_y\_agrupacion, rotulo, desc\_periodo\_terraza, desc\_situacion\_terraza, Superficie\_ES, Superficie\_RA, Fecha\_confir\_ult\_decreto\_resol, DESC\_CLASE, DESC\_NOMBRE, nom\_terraza, num\_terraza, cal\_terraza, desc\_ubicacion\_terraza, hora\_ini\_LJ\_es, hora\_fin\_LJ\_es, hora\_ini\_LJ\_ra, hora\_fin\_LJ\_ra, hora\_ini\_VS\_es, hora\_fin\_VS\_es, hora\_ini\_VS\_ra, hora\_fin\_VS\_ra, mesas\_aux\_es, mesas\_aux\_ra, mesas\_es, mesas\_ra, sillas\_es, sillas\_ra

**FROM** (

**SELECT** id\_terraza, desc\_distrito\_local, desc\_barrio\_local, clase\_vial\_edificio, desc\_vial\_edificio, nom\_edificio, num\_edificio, Cod\_Postal, coordenada\_x\_local, coordenada\_y\_local, desc\_tipo\_acceso\_local, desc\_situacion\_local, secuencial\_local\_PC, coordenada\_x\_agrupacion, coordenada\_y\_agrupacion, rotulo, desc\_periodo\_terraza, desc\_situacion\_terraza, Superficie\_ES, Superficie\_RA, Fecha\_confir\_ult\_decreto\_resol, DESC\_CLASE, DESC\_NOMBRE, nom\_terraza, num\_terraza, cal\_terraza, desc\_ubicacion\_terraza, hora\_ini\_LJ\_es, hora\_fin\_LJ\_es, hora\_ini\_LJ\_ra, hora\_fin\_LJ\_ra, hora\_ini\_VS\_es, hora\_fin\_VS\_es, hora\_ini\_VS\_ra, hora\_fin\_VS\_ra, mesas\_aux\_es, mesas\_aux\_ra, mesas\_es, mesas\_ra, sillas\_es, sillas\_ra

**FROM** "@juadaves".Datasets.Terrazas\_202104 **AS** Terrazas\_202104

) nested\_0

# Guarda la consulta con el nombre de Terreza\_001 y guárdalo en el espacio Analista 1.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated

# 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.

**SELECT** desc\_distrito\_local, id\_barrio\_local, desc\_barrio\_local, cod\_barrio\_local, id\_seccion\_censal\_local, desc\_seccion\_censal\_local, coordenada\_x\_local, coordenada\_y\_local, id\_tipo\_acceso\_local, desc\_tipo\_acceso\_local, id\_situacion\_local, desc\_situacion\_local, id\_ndp\_edificio, id\_clase\_ndp\_edificio, id\_vial\_edificio, clase\_vial\_edificio, desc\_vial\_edificio, nom\_edificio, num\_edificio, cal\_edificio, secuencial\_local\_PC, id\_ndp\_acceso, id\_clase\_ndp\_acceso, id\_vial\_acceso, clase\_vial\_acceso, desc\_vial\_acceso, nom\_acceso, num\_acceso, cal\_acceso, coordenada\_x\_agrupacion, coordenada\_y\_agrupacion, id\_agrupacion, nombre\_agrupacion, id\_tipo\_agrup, desc\_tipo\_agrup, id\_planta\_agrupado, id\_local\_agrupado, rotulo, id\_tipo\_licencia, id\_tipo\_situacion\_licencia

**FROM** "@juadaves".Datasets.Licencias\_Locales\_202104

Graphical user interface, application

Description automatically generated

Graphical user interface, application, table

Description automatically generated

# 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.

Graphical user interface, application

Description automatically generated

Graphical user interface, application, table

Description automatically generated

# Abre el dataset books y realiza los siguientes cambios:

1. Elimina el campo \_id.

Graphical user interface, text, application, email

Description automatically generated

1. Excluye los libros que no tienen ISBN (opción Exclude…, casilla *null*).

**SELECT** "\_id", isbn, pageCount, publishedDate, thumbnailUrl, shortDescription, longDescription, status, authors, categories

**FROM** (

**SELECT** "\_id", isbn, pageCount, publishedDate, thumbnailUrl, shortDescription, longDescription, status, authors, categories

**FROM** "@juadaves".Datasets.libraries

) nested\_0

**WHERE**  ( "isbn" **IS** NULL ) **IS** **FALSE**

Graphical user interface

Description automatically generated

1. 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**.

Table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Este comando extrae o desempaqueta los elementos de las listas y genera un registro nuevo de cada elemento de la lista.

# **Carga de datos de ejemplo propias de Dremio (Opcional)**

* 1. 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.

Graphical user interface, text, application, chat or text message

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

1. Crea una consulta con el nombre *Empleados* y guárdala en un espacio de trabajo llamado “*Analista General*”.

Graphical user interface, application

Description automatically generated

1. Investiga qué son los ficheros parquet e indica una diferencia (la más mencionada) con respecto a los ficheros JSON:

**Parquet:**

Apache Parquet, es un formato de almacenamiento orientado a columnas y de código libre del ecositema Apache Hadoop. Es similar a otros formatos de Hadoop como por por ejemplo RCFile y ORC. Proporciona esquemas de compresión de datos y un excelente rendimiento para manejar datos complejos.

Los valores de cada columna del archivo parquet se almacenan físicamente en memorias contiguas, este almacenamiento proporciona los siguientes beneficios:

* La compresión por columnas es eficiente y ahorra espacio de almacenamiento
* Se pueden aplicar técnicas de compresión específicas para un tipo ya que los valores de columna tienden a ser del mismo tipo
* Las consultas que obtienen valores de columna específicos no necesitan leer los datos de fila completos, lo que mejora el rendimiento
* Se pueden aplicar diferentes técnicas de codificación a diferentes columnas.

Parquet tiene una codificación automática de tipo diccionario, la cual permite una compresión significativa y aumenta la velocidad de procesamiento. Para optimizar el almacenamiento de varias apariciones del mismo valor, se almacena un solo valor una vez junto con el número de apariciones. Parquet implementa un hibrido de empaquetamiento de bits y RLE.

**Parquet vs Json:**

Json maneja un formato clave valor, mientras que el formato parquet almacena los datos en columna, el archivo Json es útil para almacenar cualquier configuración o datos tipo clave valor. El formato parquet es útil cuando almacenamos datos en formato tabular. Especialmente cuando los datos son muy grandes. Un buen ejemplo de uso de Paqruet es el escenario en el que por ejemplo tenemos datos de columnas grandes, donde el número de fila es mayor a 1.000.000, por lo general esa cantidad no se puede manipular fácilmente en un CSV. Tambien es importante mencionar que hay herramientas que nos permiten convertir un archivo en formato Json a un formato Parquet.