



Universidad Internacional de La Rioja
Facultad de Ingeniería y Tecnología

Máster Universitario en Análisis y Visualización de Datos Masivos /
Visual Analytics & 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

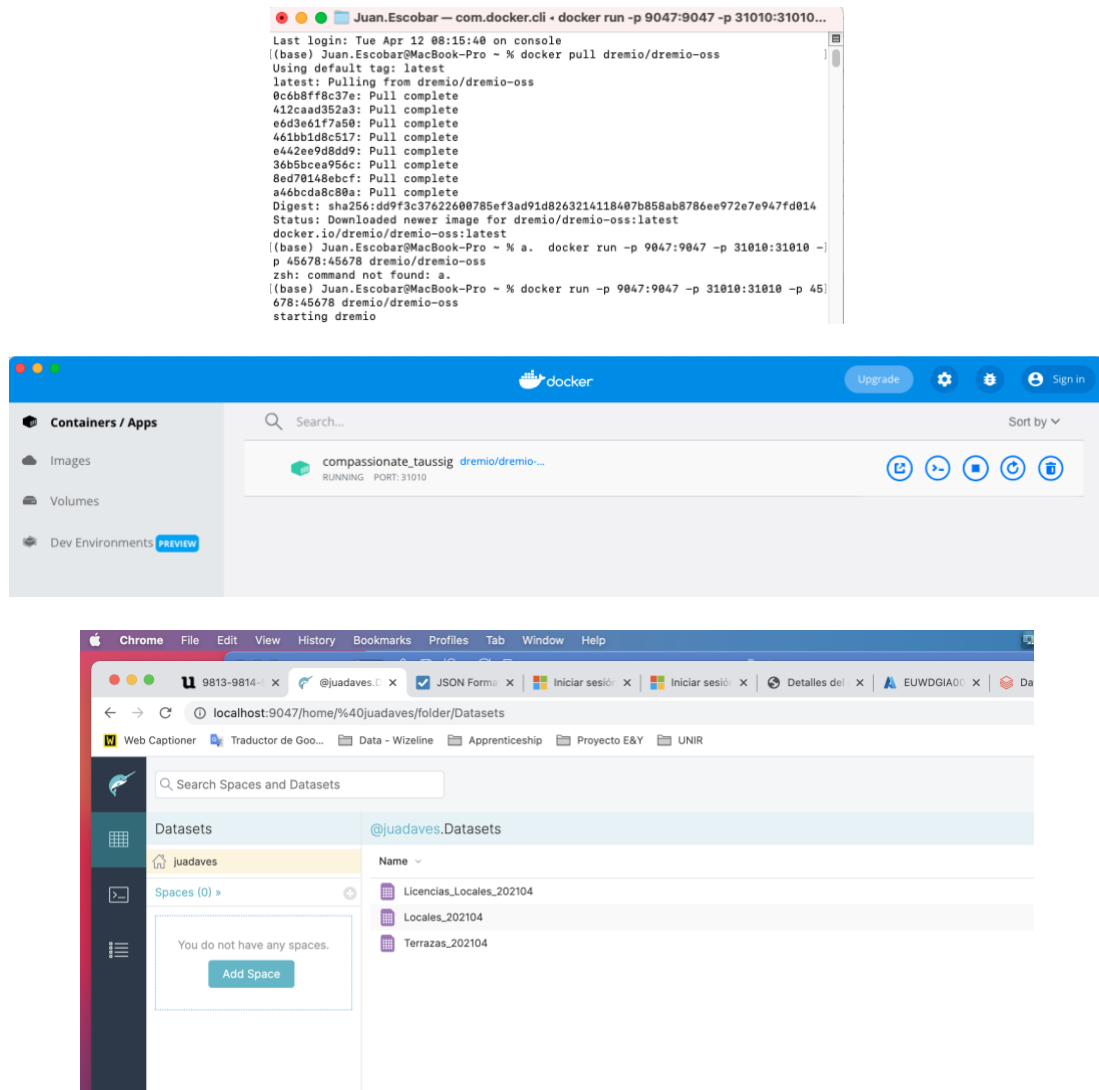
Índice de contenidos

Table of Contents

MÁSTER UNIVERSITARIO EN ANÁLISIS Y VISUALIZACIÓN DE DATOS MASIVOS / VISUAL ANALITICS & BIG DATA.....	1
1. CARGA DE DATOS.....	3
2. ESPACIOS DE TRABAJO.....	9
2.1. SOBRE LA HERRAMIENTA DEBES CREAR 3 ESPACIOS DE TRABAJOS LLAMADOS:	9
• ANALISTA 1	9
• ANALISTA 2	9
• ANALISTA 3	9
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
3. CREAR DATASETS PERSONALIZADOS	10
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
3.2. ABRE EL DATASET TERRAZAS_202104 Y REALIZA LAS SIGUIENTES MODIFICACIONES SOBRE ÉL: 10	
A. ELIMINA TODOS LOS CAMPOS ID_* EXCEPTO EL CAMPO ID_TERRAZA.	10
B. ELIMINA TODOS LOS CAMPOS ESCALERA	11
C. CREA UN NUEVO CAMPO LLAMADO SUPERFICIE_TO QUE SUME EL CAMPO SUPERFICIE_ES Y SUPERFICIE_ES.	11
D. GUARDA LA CONSULTA CON EL NOMBRE DE TERREZA_001 Y GUÁRDALO EN EL ESPACIO ANALISTA 1. 12	
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	
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
3.5. ABRE EL DATASET BOOKS Y REALIZA LOS SIGUIENTES CAMBIOS:.....	13
4. CARGA DE DATOS DE EJEMPLO PROPIAS DE DREMIO (OPCIONAL).....	15

1. Carga de datos

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



1.2. Después de instalar la herramienta, útilzala 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.

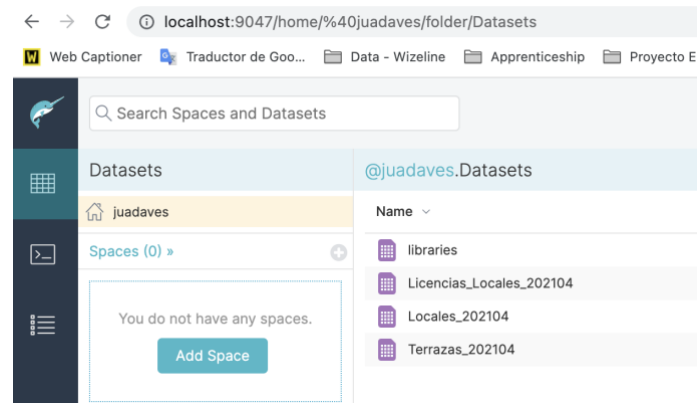
Los Dataset se suben sin problemas, excepto el Dataset libraries.json, el cual presenta un error de sintaxis en el campo_id:

```
localhost ▾ | lab ▾
1- db.libraries.insertMany([
2 { "_id" : 1, "title" : "Unlocking Android", "isbn" : "1933988673", "pageCount" : 416, "publishedDate" : { "$date": "2013-01-01T00:00:00Z" } },
3 { "_id" : 2, "title" : "Android in Action, Second Edition", "isbn" : "1935182722", "pageCount" : 592, "publishedDate" : { "$date": "2013-01-01T00:00:00Z" } },
4 { "_id" : 3, "title" : "Specification by Example", "isbn" : "1617290084", "pageCount" : 0, "publishedDate" : { "$date": "2013-01-01T00:00:00Z" } },
5 { "_id" : 4, "title" : "Flex 3 in Action", "isbn" : "1933988746", "pageCount" : 576, "publishedDate" : { "$date": "2013-01-01T00:00:00Z" } },
6 { "_id" : 5, "title" : "Flex 4 in Action", "isbn" : "1935182420", "pageCount" : 600, "publishedDate" : { "$date": "2013-01-01T00:00:00Z" } },
7 { "_id" : 6, "title" : "Collective Intelligence in Action", "isbn" : "1933988312", "pageCount" : 425, "publishedDate" : { "$date": "2013-01-01T00:00:00Z" } },
8 { "_id" : 7, "title" : "Zend Framework in Action", "isbn" : "1933988320", "pageCount" : 432, "publishedDate" : { "$date": "2013-01-01T00:00:00Z" } },
9 { "_id" : 8, "title" : "Flex on Java", "isbn" : "1933988797", "pageCount" : 265, "publishedDate" : { "$date": "2013-01-01T00:00:00Z" } },
10 ]
0.056 s
1- bulkWriteError({
2   "writeErrors" : [
3     {
4       "index" : 399,
5       "code" : 52,
6       "errmsg" : "_id fields may not contain '$'-prefixed fields: $oid is not valid for storage.",
7       "op" : {
8         "_id" : {
9           "$oid" : "53c2ae8528d75d572c06ad9d"
10         }
11       }
12     }
13   ]
14 })
```

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.

```
18 |" : /95, "title": "Learn Git in a Month of Lunches",
19 |" : 796, "title": "Understanding SPAs", "isbn": "1617
20 |" : { "oid": "53c2ae8528d75d572c06ad9d" }, "title":
21 |" : { "oid": "53c2ae8528d75d572c06ad9e" }, "title":
22 |" : { "oid": "53c2ae8528d75d572c06ad9f" }, "title":
23 |" : { "oid": "53c2ae8528d75d572c06ada0" }, "title":
24 |" : { "oid": "53c2ae8528d75d572c06ada1" }, "title":
25 |" : { "oid": "53c2ae8528d75d572c06ada2" }, "title":
```

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



1.3. 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.

Descripción metadatos tabal Terrazas_202104

Terrazas

1. id_terraza : código que identifica la terraza
2. id_local : "Código numérico que identifica cada local. Además cada local queda identificado por el código del NDP del edificio al que pertenece más el secuencial de local si es un puerta de calle o por el código de la agrupación mas la planta y el local si es un local agrupado."
3. id_distrito_local : Código numérico con el distrito municipal
4. desc_distrito_local : Literal del distrito municipal
5. id_barrio_local : "Código numérico con del barrio municipal (incluye el código de distrito)"
6. desc_barrio_local : Literal del barrio municipal
7. id_ndp_edificio : "Código numérico que identifica la dirección principal del edificio en el que se ubica el local (tipo de vía, nombre de vía, nominal, número y calificador). Por ejemplo "C/Goya num 24" tiene asociado el código de NDP "11014430". "
8. id_clase_ndp_edificio : "Código numérico que identifica el tipo de dirección: "1" si es una dirección normalizada (oficial) y "9" si es una dirección no normalizada. Para los tipos de acceso mostrados no debe de aparecer ningún clase "9:"
9. id_vial_edificio : "Código numérico que identifica la clase más el nombre de vía del edificio "
10. clase_vial_edificio : Recoge si se trata de Calle, Avenida, Plaza...
11. desc_vial_edificio : Nombre de la vía según el callejero oficial
12. nom_edificio : Tipo de numeración (Número, Kilómetro, bloque...)
13. num_edificio : Número de la calle
14. Cod_Postal : Código postal
15. coordenada_x_local : "Coordenadas UTM que identifican, de forma aproximada, la entrada principal al local puerta de calle (Sistema de referencia: Hasta el 15 de septiembre de 2017 ED-50, a partir de esa fecha ETRS89). Sólo disponen de esta información los locales tipo de acceso "Puerta de Calle" que no estén en situación de "Baja", los agrupados tienen información de las coordenadas de la agrupación (ver campos coordenadas_x/y_agrupación y los interiores no disponen de coordenadas) "
16. coordenada_y_local : "Coordenadas UTM que identifican, de forma aproximada, la entrada principal al local puerta de calle (Sistema de referencia: Hasta el 15 de septiembre de 2017 ED-50, a partir de esa fecha ETRS89). Sólo disponen de esta información los locales tipo de acceso "Puerta de Calle" que no estén en situación de "Baja", los agrupados tienen información de las coordenadas de la agrupación (ver campos coordenadas_x/y_agrupación y los interiores no disponen de coordenadas) "
17. id_tipo_acceso_local : Código numérico que identifica el tipo de acceso (ver tabla "Tipo acceso")
18. desc_tipo_acceso_local : Tipos de local según su acceso (ver tabla "Tipo acceso")
19. id_situacion_local : "Código numérico que identifica la situación de un local (ver tabla "Situación") "
20. desc_situacion_local : Tipos de situación de un local (ver tabla "Situación")
21. secuencial_local_PC : "Número secuencial que comienza en 10 en cada edificio y se asigna en saltos decenales. Identifica cada local puerta de calle de cada edificio y la asignación se realiza empezando por el primer local de la izquierda situado en la fachada del edificio donde está el portal principal, recorriendo el edificio completo en sentido contrario a las agujas del reloj. Este número junto con el código de ndp constituye el código identificativo de local."
22. Escalera : Información de la escalera del edificio que sólo aparecerá rellena, en caso de que exista, para locales con tipo de acceso "Interior"
23. id_planta_agrupado : Información relativa a la planta que sólo aparecerá rellena para los locales con tipo de acceso "Agrupado" o "Interior"
24. id_local_agrupado : Información relativa a la puerta que sólo aparecerá rellena para los locales con tipo de acceso "Agrupado" o "Interior" (en el caso de interiores esta información se corresponde con la "puerta")
25. coordenada_x_agrupacion : Coordenadas UTM que identifican, de forma aproximada, la entrada principal a la agrupación
26. coordenada_y_agrupacion : Coordenadas UTM que identifican, de forma aproximada, la entrada principal a la agrupación
27. rotulo : Nombre comercial del establecimiento
28. id_periodo_terraza : Informa acerca del periodo de funcionamiento de la terraza; puede ser '1' Anual y '2' Estacional
29. desc_periodo_terraza : Literal asociado al valor del campo anterior: 'Anual' y 'Estacional'
30. id_situacion_terraza : Informa de si la terraza consta en Censo de Locales como '1' (abierta) o '8' "Suspensión temporal"
31. desc_situacion_terraza : Informa de si la terraza consta en Censo de Locales como 'Abierta' o como "Suspensión temporal"
32. Superficie_ES : superficie en metros cuadrados ocupada por la terraza en periodo 'estacional' (Nota: La ordenanza contempla que las terrazas anuales tengan diferentes superficies de ocupación durante el periodo estacional respecto al resto del año)
33. Superficie_RA : superficie en metros cuadrados ocupada por la terraza en periodo 'resto del año'
34. Fecha_confir_ult_decreto_resol : "Fecha en la que se confirma el decreto de resolución del expediente correspondiente (Renovación, Cambio




Wiki


Descripción metadatos Locales_202104


Locales


1. **id_local**: "Código numérico que identifica cada local. Además cada local queda identificado por el código del NDP del edificio al que pertenece más el secuencial de local si es un puerta de calle o por el código de la agrupación mas la planta y el local si es un local agrupado. "
2. **id_distrito_local**: Código numérico con el distrito municipal
3. **desc_distrito_local**: Literal del distrito municipal
4. **id_barrio_local**: "Código numérico con del barrio municipal (incluye el código de distrito) "
5. **desc_barrio_local**: Literal del barrio municipal
6. **cod_barrio_local**: "Código numérico con el barrio municipal (Se requiere el dato de distrito para identificar de forma inequívoca el barrio) "
7. **id_seccion_censal_local**: Código de distrito más sección censal
8. **desc_seccion_censal_local**: "Código de sección censal. (Se requiere el dato de distrito para identificar de forma inequívoca a cada sección) "
9. **coordenada_x_local**: "Coordenadas UTM que identifican, de forma aproximada, la entrada principal al local puerta de calle (Sistema de referencia: Hasta el 15 de septiembre de 2017 ED-50, a partir de esa fecha ETRS89). Sólo disponen de esta información los locales tipo de acceso "Puerta de Calle" que no estén en situación de "Baja", los agrupados tienen información de las coordenadas de la agrupación (ver campos coordenadas_x/y_agrupación y los interiores no disponen de coordenadas) "
10. **coordenada_y_local**: "Coordenadas UTM que identifican, de forma aproximada, la entrada principal al local puerta de calle (Sistema de referencia: Hasta el 15 de septiembre de 2017 ED-50, a partir de esa fecha ETRS89). Sólo disponen de esta información los locales tipo de acceso "Puerta de Calle" que no estén en situación de "Baja", los agrupados tienen información de las coordenadas de la agrupación (ver campos coordenadas_x/y_agrupación y los interiores no disponen de coordenadas) "
11. **id_tipo_acceso_local**: "Código numérico que identifica el tipo de acceso (ver tabla "Tipo acceso") "
12. **desc_tipo_acceso_local**: Tipos de local según su acceso
13. **id_situacion_local**: "Código numérico que identifica la situación de un local (ver tabla "Situación") 0 Agrupado Local perteneciente a una agrupación
1 Puerta calle Local con acceso desde la calle
12 Puerta de calle asociado Permite asociar a un Puerta de Calle varios titulares con actividades diferente"
14. **desc_situacion_local**: Tipos de local según su acceso
15. **id_vial_edificio**: "Código numérico que identifica la clase más el nombre de vía del edificio "
16. **clase_vial_edificio**: Recoge si se trata de Calle, Avenida, Plaza...
17. **desc_vial_edificio**: Nombre de la vía según el callejero oficial
18. **id_ndp_edificio**: "Código numérico que identifica una dirección (tipo de vía, nombre de vía, nominal, número y calificador). Por ejemplo "C/Goya num 24" tiene asociado el código de NDP "11014430". "
19. **id_clase_ndp_edificio**: " Código numérico que identifica el tipo de dirección: "1" si es una dirección normalizada (oficial) y "9" si es una dirección no normalizada. Para los tipos de acceso mostrados no debe de aparecer ningun clase "9"."
20. **nom_edificio**: Tipo de numeración (Número, Kilómetro, bloque...)



Licencias_Locales_202104
@juadaves.Datasets


Data

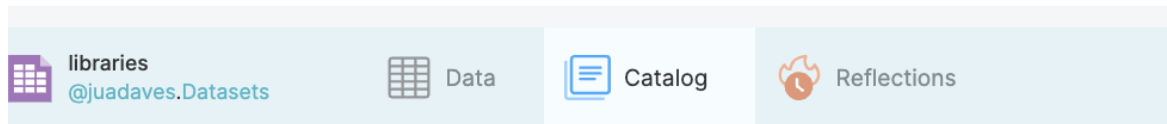

Catalog


Reflections

Descripción metadata Licencias_locales_202104

Licencias

1. **id_local**: "Código numérico que identifica cada local. Además cada local queda identificado por el código del NDP del edificio al que pertenece más el secuencial de local si es un puerta de calle o por el código de la agrupación mas la planta y el local si es un local agrupado."
2. **id_distrito_local**: Código numérico con el distrito municipal
3. **desc_distrito_local**: Literal del distrito municipal
4. **id_barrio_local**: Código numérico con del barrio municipal (incluye el código de distrito)
5. **desc_barrio_local**: Literal del barrio municipal
6. **cod_barrio_local**: "Código numérico con el barrio municipal (Se requiere el dato de distrito para identificar de forma inequívoca el barrio) "
7. **id_seccion_censal_local**: Código de distrito más sección censal
8. **desc_seccion_censal_local**: "Código de sección censal. (Se requiere el dato de distrito para identificar de forma inequívoca a cada sección) "
9. **coordenada_x_local**: " Coordenadas UTM que identifican, de forma aproximada, la entrada principal al local puerta de calle (Sistema de referencia: Hasta el 15 de septiembre de 2017 ED-50, a partir de esa fecha ETRS89). Sólo disponen de esta información los locales tipo de acceso "Puerta de Calle" que no estén en situación de "Baja", los agrupados tienen información de las coordenadas de la agrupación (ver campos coordenadas_x/y_agrupación y los interiores no disponen de coordenadas) "
10. **coordenada_y_local**: " Coordenadas UTM que identifican, de forma aproximada, la entrada principal al local puerta de calle (Sistema de referencia: Hasta el 15 de septiembre de 2017 ED-50, a partir de esa fecha ETRS89). Sólo disponen de esta información los locales tipo de acceso "Puerta de Calle" que no estén en situación de "Baja", los agrupados tienen información de las coordenadas de la agrupación (ver campos coordenadas_x/y_agrupación y los interiores no disponen de coordenadas) "
11. **id_tipo_acceso_local**: "Código numérico que identifica el tipo de acceso (ver tabla "Tipo acceso") "
12. **desc_tipo_acceso_local**: Tipos de local según su acceso
13. **id_situacion_local**: "Código numérico que identifica la situación de un local en relación con la actividad (ver tabla "Situación"). Indicar que se trata de una variable de mantenimiento complicado, ya que, aunque ya están funcionando diferentes procedimientos de actualización de los datos de actividad, no se dispone de ninguno que informe de cuándo una actividad cesa y el local se cierra sin que aparezca una nueva actividad. "
14. **desc_situacion_local**: Tipos de situación de un local
15. **id_ndp_edificio**: "Código numérico que identifica una dirección (tipo de vía, nombre de vía, nominal, número y calificador). Por ejemplo "C/Goya num 24" tiene asociado el código de NDP "11014430". "
16. **id_clase_ndp_edificio**: "Código numérico que identifica el tipo de dirección: "1" si es una dirección normalizada (oficial) y "9" si es una dirección no normalizada. Para los tipos de acceso mostrados no debe de aparecer ningún clase "9"."
17. **id_vial_edificio**: "Código numérico que identifica la clase más el nombre de vía del edificio "
18. **clase_vial_edificio**: Recoge si se trata de Calle, Avenida, Plaza...
19. **desc_vial_edificio**: Nombre de la vía según el callejero oficial
20. **nom_edificio**: Tipo de numeración (Número, Kilómetro, bloque...)
21. **num_edificio**: Número de la calle



Wiki

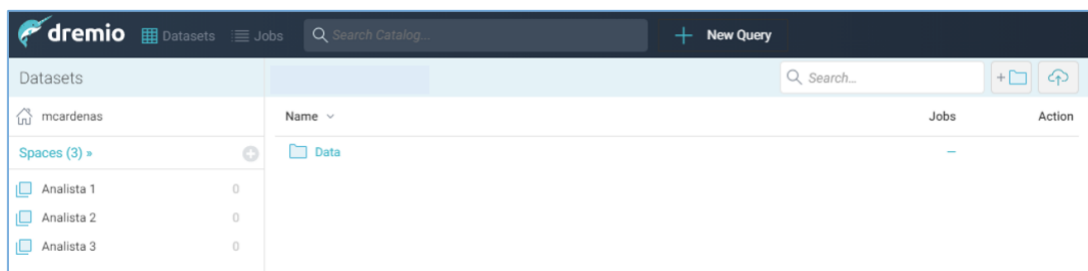
Descripción de los metadatos de la tabla libraries:

1. **_id**: Identificador del libro, tipo numerico consecutivo
2. **title**: Titulo del libro
3. **isbn**: Código normalizado internacional para libros (desde enero de 2007, tienen siempre 13 dígitos)
4. **pageCount**: Cantidad de paginas del libro
5. **publishedDate**: Fecha de publicación
6. **thumbnailUrl**: URL thumbnail imagen miniatura del libro
7. **shortDescription**: Descripción corta del libro
8. **longDescription**: Descripción larga del libro
9. **status**: Estado del libreo (PUBLISH or MEAP)
10. **authors**: Lista de autores del libro
11. **categories**: Losta de categorias o tematicas del libro

2. Espacios de trabajo

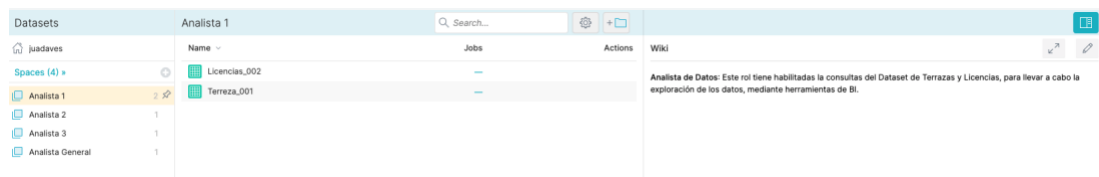
2.1. Sobre la herramienta debes crear 3 espacios de trabajos llamados:

- Analista 1
- Analista 2
- Analista 3

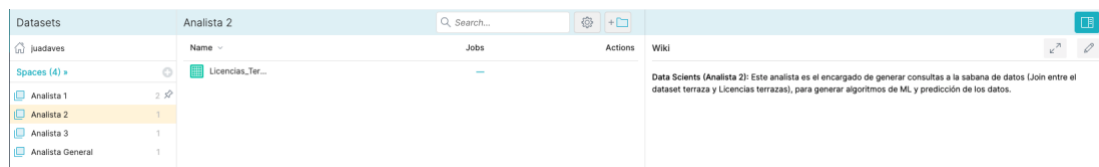


2.2. Cada espacio de trabajo debe llevar una “wiki content” que explique la finalidad del espacio de trabajo. Esto es una información libre que, también debe describir qué contiene el espacio de trabajo. Por ejemplo:

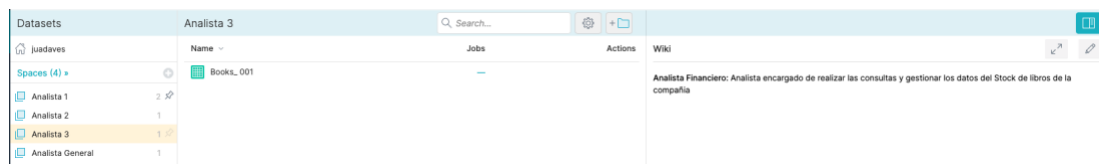
- **Analista 1**: agrupa datos relacionados con los ficheros CSV. Estos ficheros tienen que ver con la información de locales de la ciudad.



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

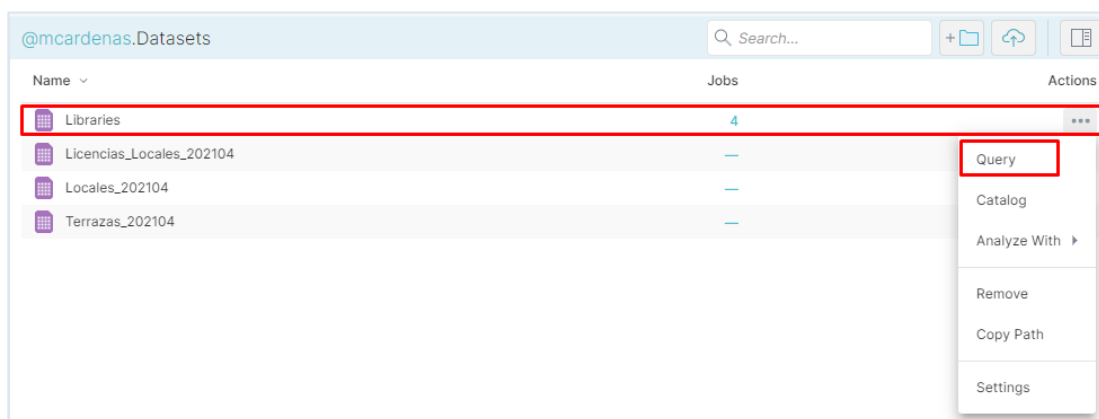


- **Analista 3:** agrupa datos relacionados con los ficheros Open Data. Estos ficheros tienen información sobre el clima de la ciudad.



3. Crear datasets personalizados

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.

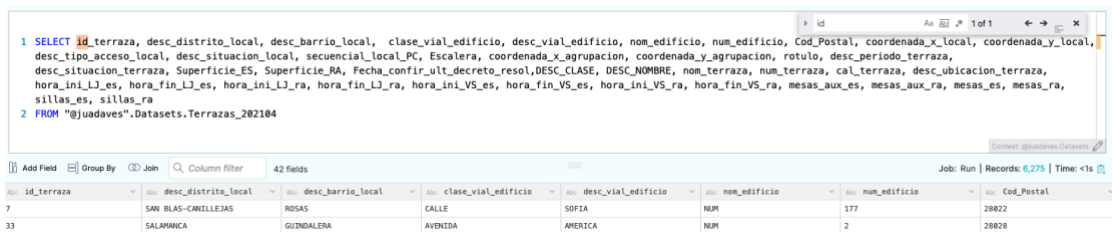


3.2. Abre el dataset Terrazas_202104 y realiza las siguientes modificaciones sobre él:

- Elimina todos los campos id_* excepto el campo id_terraza.

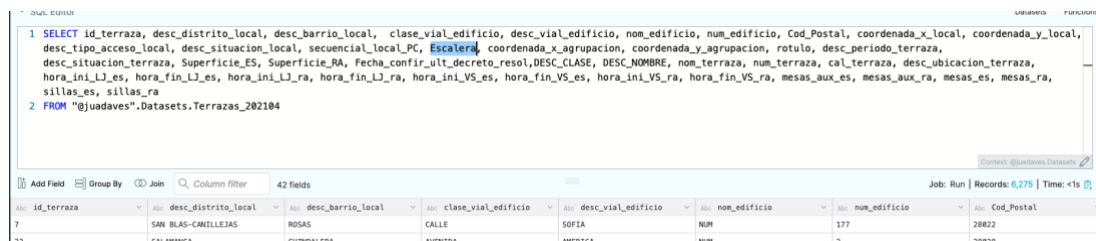
```
SELECT id_terrazza, desc_distrito_local,
desc_barrio_local, clase_vial_edificio, desc_vial_edificio, nom_edificio,
num_edificio, Cod_Postal, coordenada_x_local,
coordenada_y_local, desc_tipo_acceso_local, desc_situacion_local,
secuencial_local_PC, Escalera, coordenada_x_agrupacion,
coordenada_y_agrupacion, rotulo,
desc_periodo_terrazza, desc_situacion_terrazza, Superficie_ES, Superficie_RA,
Fecha_confir_ult_decreto_resol, DESC_CLASE, DESC_NOMBRE, nom_terrazza,
num_terrazza, cal_terrazza, desc_ubicacion_terrazza, 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
```



id_terrazza	desc_distrito_local	desc_barrio_local	clase_vial_edificio	desc_vial_edificio	nom_edificio	num_edificio	Cod_Postal
7	SAN BLAS-CANILLEJAS	ROSAS	CALLE	SOFA	NUM	177	28822
33	SALAMANCA	GUINDALERA	AVENIDA	AMERICA	NUM	2	28828

b. Elimina todos los campos Escalera



id_terrazza	desc_distrito_local	desc_barrio_local	clase_vial_edificio	desc_vial_edificio	nom_edificio	num_edificio	Cod_Postal
7	SAN BLAS-CANILLEJAS	ROSAS	CALLE	SOFA	NUM	177	28822
33	SALAMANCA	GUINDALERA	AVENIDA	AMERICA	NUM	2	28828



id_terrazza	desc_distrito_local	desc_barrio_local	clase_vial_edificio	desc_vial_edificio	nom_edificio	num_edificio	Cod_Postal
13	SAN BLAS-CANILLEJAS	ROSAS	CALLE	SOFA	NUM	177	28822
14	SALAMANCA	GUINDALERA	CALLE	ALONSO HEREDIA	NUM	25	28828

c. Crea un nuevo campo llamado Superficie_TO que sume el campo Superficie_ES y Superficie_ES.

```
SELECT id_terrazza, 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_terrazza, desc_situacion_terrazza,
Superficie_ES, Superficie_RA, Fecha_confir_ult_decreto_resol, DESC_CLASE,
DESC_NOMBRE, nom_terrazza, num_terrazza, cal_terrazza, desc_ubicacion_terrazza,
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 (
```

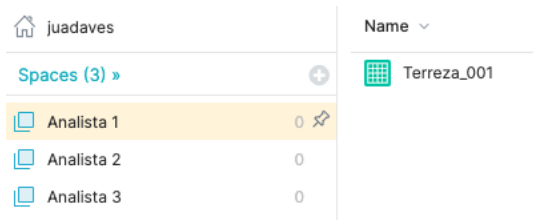
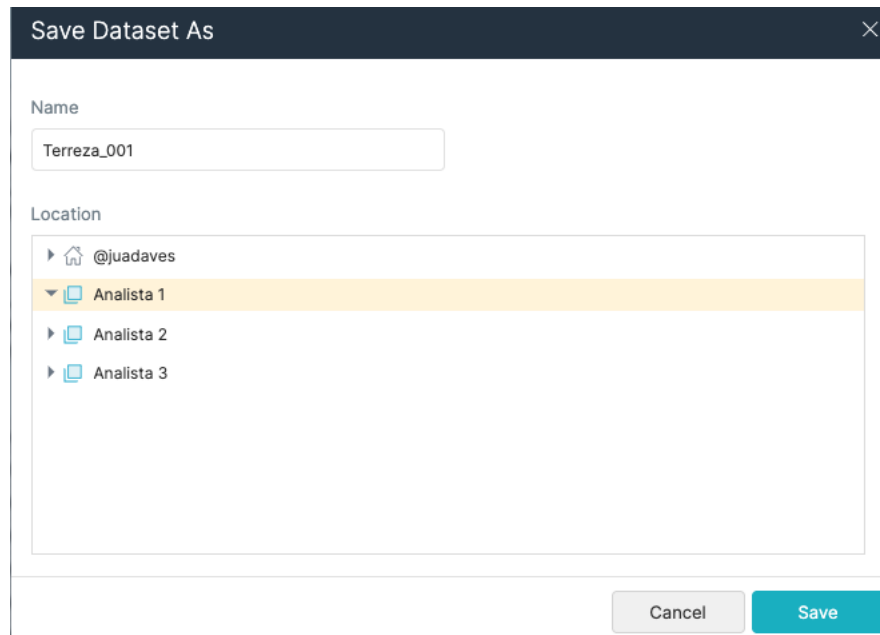
```
SELECT id_terrazza, desc_distrito_local, desc_barrio_local,
clase_vial_edificio, desc_vial_edificio, nom_edificio, num_edificio,
Cod_Postal, coordenada_x_local, coordenada_y_local, desc_tipo_acceso_local,
```

```
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
```

- d. Guarda la consulta con el nombre de Terreza_001 y guárdalo en el espacio Analista 1.



- 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.

```
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
```

Save Dataset As

Name

Licencias_002

Location

- ▶ @juadaves
- ▶ Analista 1
- ▶ Analista 2
- ▶ Analista 3

Datasets		Analista 2
juadaves		Name ▾
Spaces (3) »		Licencias_002
Analista 1	1	
Analista 2	1	
Analista 3	0	

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.

Recommended Join Custom Join

Type: Inner

Select fields from "Terrazas_202104" (c...

id_local

Select fields from "Licencias_Locales_2..."

id_local

Apply Preview Cancel ⚠ Result based on sample dataset

id_terrazas	id_local	id_distrito_local	desc_distrito_local	id_barrio_local	desc_barrio_local	id_ndp_edificio	id_clase_ndp_edificio
7	288867128	28	SAN BLAS-CRISTLEJAS	2885	RDSAS	28157611	1
13	176483156	4	CAL ABASO'S	AB4	CUTIBANAI EDA	17495878	1

Datasets		Analista 2
juadaves		Name ▾
Spaces (3) »		Licencias_Terrazas_003
Analista 1	2	
Analista 2	2	
Analista 3	0	

3.5. Abre el dataset books y realiza los siguientes cambios:

a. Elimina el campo id.

SQL Editor

```
1 SELECT isbn, pageCount, publishedDate, thumbnailUrl, shortDescription, longDescription, status, authors, categories
2 FROM "juadaves".Datasets.Libraries
```

Add Field Group By Join Column filter 9 fields

isbn	pageCount	publishedDate	thumbnailUrl	shortDescription	longDescription	status	authors
1933988673	416	2009-04-01 00:00:00.000	https://s3.amazonaws.com/AKIA Unlocking Android: A Developer's Guide to Android	Unlocking Android: A Developer's Guide to Android	Android is an open source mob	PUBLISH	["W. Frank Ableson","Charlie
1935182722	592	2011-01-14 00:00:00.000	https://s3.amazonaws.com/AKIA Android in Action, Second Edition	Android in Action, Second Edition	When it comes to mobile apps. PUBLISH	PUBLISH	["W. Frank Ableson","Robi Ser

Job: Run | Records: 431 | Time: <1s

b. Excluye los libros que no tienen ISBN (opción Exclude..., casilla null).

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

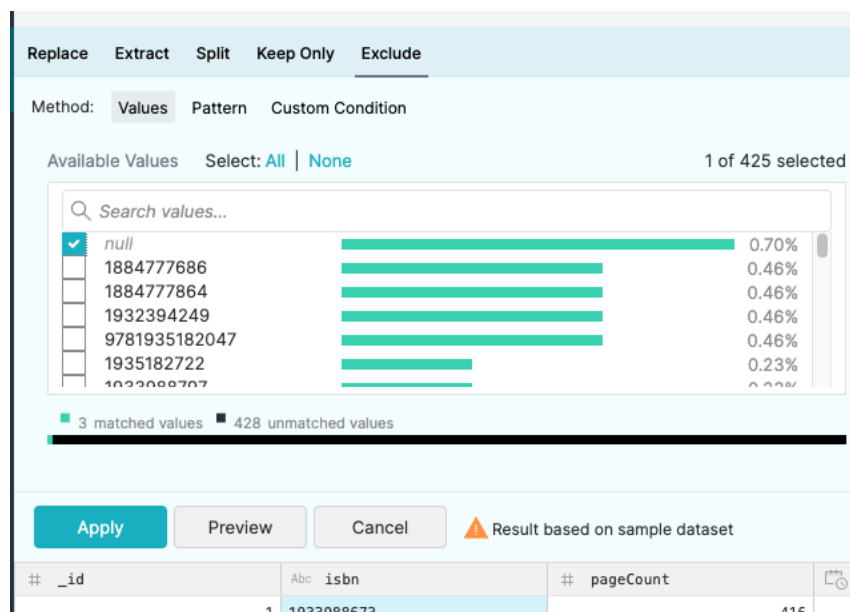
```
FROM (
```

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

```
FROM "@juadaves".Datasets.libraries
```

```
) nested_0
```

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



- c. 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**.

#	_id	Abc authors	Abc categories
	1	W. Frank Ableson	Open Source
	1	W. Frank Ableson	Mobile
	1	Charlie Collins	Open Source
	1	Charlie Collins	Mobile
	1	Robi Sen	Open Source
	1	Robi Sen	Mobile
	2	W. Frank Ableson	Java
	2	Robi Sen	Java
	3	Gojko Adzic	Software Engineering
	4	Tariq Ahmed with Jor	Internet
	4	Faisal Abid	Internet
	5	Tariq Ahmed	Internet
	5	Dan Orlando	Internet
	5	John C. Bland II	Internet
	5	Joel Hooks	Internet
	6	Satnam Alag	Internet
	7	Rob Allen	Web Development

Save Dataset As

Name

Books_001|

Location

- ▶ @juadaves
- ▶ Analista 1
- ▶ Analista 2
- ▶ Analista 3

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

4. Carga de datos de ejemplo propias de Dremio (Opcional)

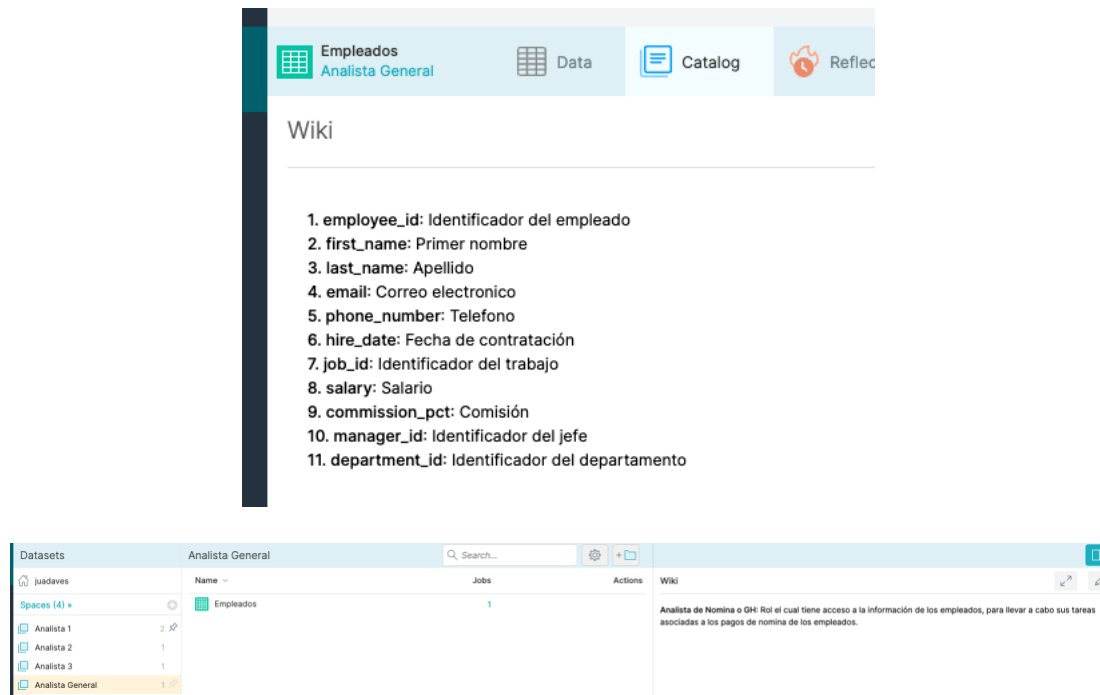
4.1. Desde el repositorio de datos de ejemplo de Dremio, crea un espacio de trabajo con dichos datos, en concreto con el fichero *employees.parquet*.

- a) Crea una Wiki Content para este nuevo repositorio.

Samples.samples.dremio.com.Dremio University

Name

- 4week_recipes.json
- 100_Sales_Records_inconsistency.csv
- aac_shelter_outcomes.csv
- airbnb_listings.csv
- employees.parquet



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

Datasets	Analista General
juadaves	Name
Spaces (4) »	Empleados
Analista 1	2
Analista 2	1
Analista 3	1
Analista General	1

- c) 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 ecosistema Apache Hadoop. Es similar a otros formatos de Hadoop como 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 híbrido 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 Parquet 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. También es importante mencionar que hay herramientas que nos permiten convertir un archivo en formato Json a un formato Parquet.