



CURSO DE ESPECIALIZACIÓN EN INTELIGENCIA ARTIFICIAL Y BIGDATA

BIGDATA APLICADO

## TAREA EVALUABLE 4.1

**Autor:** Carlos Sánchez Recio.  
06 / 01 / 2025

# Índice

<b>Apartado 3: Centros educativos de las Islas Baleares</b>	<b>1</b>
1      Recupera el número de centros públicos ( <i>esPublic</i> será true) de la isla de Eivissa. . . . .	3
2      Recupera el nombre de todos los institutos de educación secundaria del municipio ( <i>nomMunicipi</i> ) de Palma. . . . .	3
3      Recupera el número de centros de cada tipo ( <i>tipusCentreNomCa</i> ) de la isla de Menorca. . . . .	4
4      Recupera el nombre de todos los centros de la isla de Mallorca que ofrecen estudios de la etapa ( <i>tipusCentreNomCa</i> ) "Grau superior". . . . .	4
<b>Actualización 09/01/2025</b>	<b>6</b>

## Apartado 3: Centros educativos de las Islas Baleares

En el [catálogo de datos abiertos de las Islas Baleares](#) podemos encontrar un dataset con los [centros educativos de las Islas Baleares](#), en formato JSON. También puedes encontrarlo en el repositorio del curso. Descarga este archivo y sube a un directorio de HDFS.

Crea una tabla en el almacén de datos de Hive, de modo que se pueda utilizar en Impala, y carga los datos que tenemos en el archivo JSON. Las etapas educativas ( nom Etapa ) deben tratarse como un tipo complejo ARRAY.

**Importante:** No puedes editar el archivo previamente, debe cargarse tal cual está publicado.

### ¡ATENCIÓN!

¡El día 3/1/2025 han borrado los datos de todos los centros educativos del catálogo de datos abiertos!

Puede trabajar con una copia del JSON correcto: [centres\\_educatius.json](#) (aunque sólo contiene los 100 primeros centros).

El primer paso es obtener el archivo en el servidor y subirlo a Hadoop mediante la siguiente secuencia de comandos:

```
1 wget https://raw.githubusercontent.com/tnavarrete-iedib/bigdata-24-25/refs
2   /heads/main/centres_educatius2.json
3 hdfs dfs -mkdir centres
4 # Opcional:
5 mv centres_educatius2.json centres2.json
6 hdfs dfs -put centres2.json centres/
```

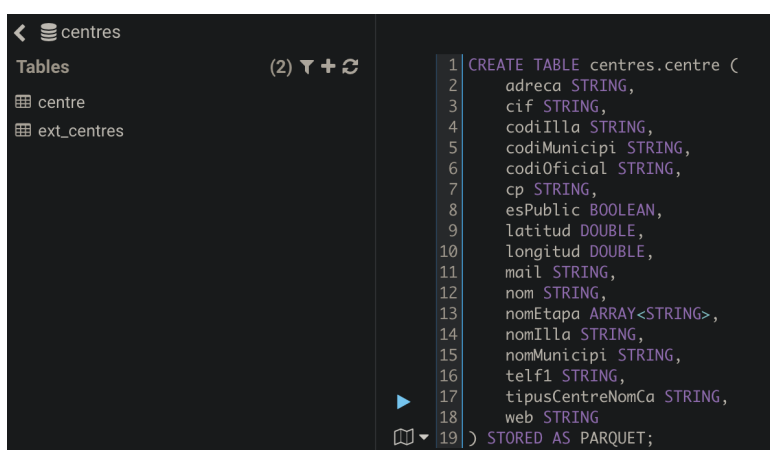
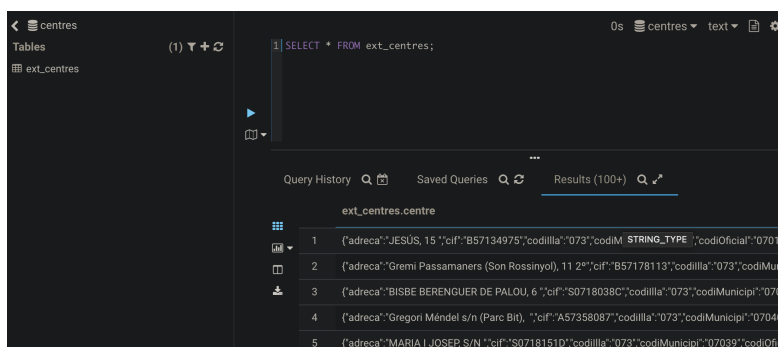
El siguiente paso es acceder a la interfaz HUE en el editor de Hive y ejecutar las siguientes sentencias para crear la base de datos y tabla además de cargar los datos en ella:

```
1 CREATE DATABASE centres;
2
3 USE centres;
4
5 CREATE EXTERNAL TABLE ext_centres (centre STRING)
6 LOCATION '/user/cloudera/centres';
7
8 SELECT * FROM ext_centres; -- Test
9
10 CREATE TABLE centres.centre (
11     adreca STRING,
12     cif STRING,
13     codiIlla STRING,
14     codiMunicipi STRING,
15     codiOficial STRING,
16     cp STRING,
17     esPublic BOOLEAN,
18     latitud DOUBLE,
19     longitud DOUBLE,
20     mail STRING,
21     nom STRING,
22     nomEtapa ARRAY<STRING>,
23     nomIlla STRING,
24     nomMunicipi STRING,
25     telf1 STRING,
26     tipusCentreNomCa STRING,
```

```

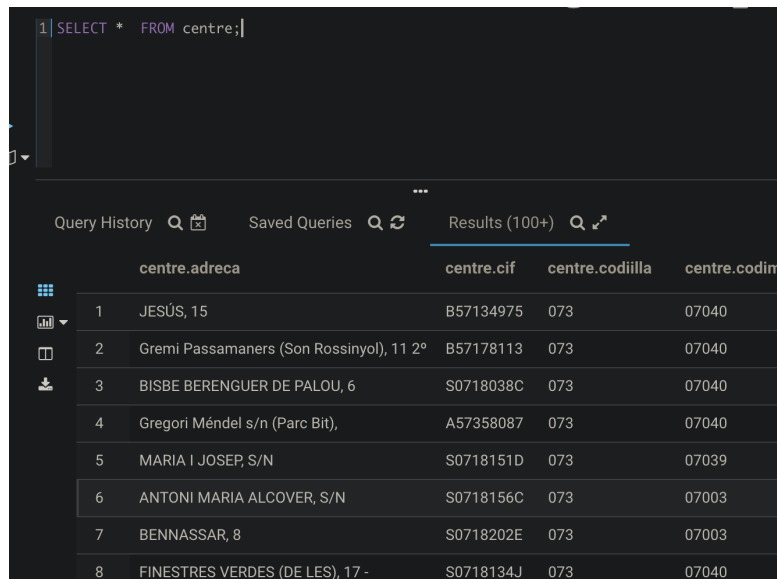
27     web STRING
28 ) STORED AS PARQUET;
29
30 INSERT INTO TABLE centre
31 SELECT
32     get_json_object(centre, '$.adreca') AS adreca,
33     get_json_object(centre, '$.cif') AS cif,
34     get_json_object(centre, '$.codiIlla') AS codiIlla,
35     get_json_object(centre, '$.codiMunicipi') AS codiMunicipi,
36     get_json_object(centre, '$.codiMunicipi') AS codiMunicipi,
37     get_json_object(centre, '$.cp') AS cp,
38     CAST(get_json_object(centre, '$.esPublic') AS BOOLEAN) AS esPublic,
39     CAST(get_json_object(centre, '$.latitud') AS DOUBLE) AS latitud,
40     CAST(get_json_object(centre, '$.longituf') AS DOUBLE) AS longitud,
41     get_json_object(centre, '$.mail') AS mail,
42     get_json_object(centre, '$.nom') AS nom,
43     SPLIT(get_json_object(centre, '$.nomEtapas'), ',') AS nomEtapas,
44     get_json_object(centre, '$.nomIlla') AS nomIlla,
45     get_json_object(centre, '$.nomMunicipi') AS nomMunicipi,
46     get_json_object(centre, '$.telf1') AS telf1,
47     get_json_object(centre, '$.tipusCentreNomCa') AS tipusCentreNomCa,
48     get_json_object(centre, '$.web') AS web
49 FROM ext_centres;
50
51 SELECT * FROM centre; -- Test

```



Tras finalizar estos pasos, antes de empezar a ejecutar sentencias en el editor de Impala, es necesario ejecutar la siguiente sentencia en el editor de Impala:

```
1 INVALIDATE METADATA centres.centre;
```



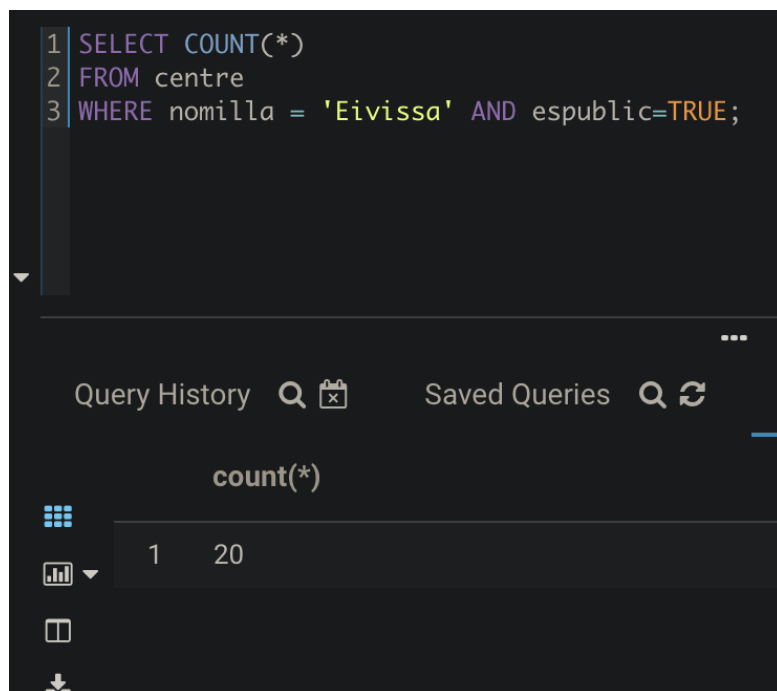
The screenshot shows a database query interface. At the top, a query editor contains the text: `1|SELECT * FROM centre;|`. Below the editor, there are tabs for 'Query History', 'Saved Queries', and 'Results (100+)'. The 'Results' tab is active, displaying a table with 8 rows and 4 columns: `centre.adreca`, `centre.cif`, `centre.codilla`, and `centre.codim`.

	centre.adreca	centre.cif	centre.codilla	centre.codim
1	JESÚS, 15	B57134975	073	07040
2	Gremi Passamaners (Son Rossinyol), 11 2º	B57178113	073	07040
3	BISBE BERENGUER DE PALOU, 6	S0718038C	073	07040
4	Gregori Méndel s/n (Parc Bit),	A57358087	073	07040
5	MARIA I JOSEP, S/N	S0718151D	073	07039
6	ANTONI MARIA ALCOVER, S/N	S0718156C	073	07003
7	BENNASSAR, 8	S0718202E	073	07003
8	FINESTRES VERDES (DE LES), 17 -	S0718134J	073	07040

A continuación, utilizando Impala (ya sea desde Hue o desde el shell), debes ejecutar las siguientes consultas:

1 Recupera el número de centros públicos (*esPublic* será true) de la isla de Eivissa.

```
1 SELECT COUNT(*)
2 FROM centre
3 WHERE nomilla = 'Eivissa' AND espublic=TRUE;
```



The screenshot shows a database query interface. At the top, a query editor contains the text: `1|SELECT COUNT(*)
2|FROM centre
3|WHERE nomilla = 'Eivissa' AND espublic=TRUE;`. Below the editor, there are tabs for 'Query History' and 'Saved Queries'. The 'Results' tab is active, displaying a table with 1 column: `count(*)`. The result is 20.

count(*)
20

2 Recupera el nombre de todos los institutos de educación secundaria del municipio (*nomMunicipi*) de Palma.

```
1 SELECT nom
2 FROM centre
3 WHERE nommunicipi='Palma';
```

```
4 | AND LOWER(tipuscentrenomca) LIKE '%institut%educaci secund ria%';
```

```
1 SELECT nom
2 FROM centre
3 WHERE nommunicipi='Palma'
4 AND LOWER(tipuscentrenomca) LIKE '%institut%educació secundària%';
```

Query History Saved Queries Results (2)

	nom	STRING_TYPE
1	ANTONI MAURA	
2	ARXIDUC LLUÍS SALVADOR	

3 Recupera el número de centros de cada tipo (*tipusCentreNomCa*) de la isla de Menorca.

```
1 SELECT COUNT(tipuscentrenomca), tipuscentrenomca
2 FROM centre
3 WHERE nomilla='Menorca'
4 GROUP BY tipuscentrenomca;
```

```
1 SELECT COUNT(tipuscentrenomca), tipuscentrenomca
2 FROM centre
3 WHERE nomilla='Menorca'
4 GROUP BY tipuscentrenomca;
```

Query History Saved Queries Results (5)

	count(tipuscentrenomca)	tipuscentrenomca
1	1	Institut d'educació secundària
2	1	Escola infantil (entitats locals)
3	1	Centre privat d'educació especial
4	2	Col·legi d'educació infantil i primària
5	7	Centre d'educació de persones adultes

4 Recupera el nombre de todos los centros de la isla de Mallorca que ofrecen estudios de la etapa ( *tipusCentreNomCa* ) "Grau superior".

```
1 SELECT nom
2 FROM centres.centre t, t.nomEtapa etapa
3 WHERE nomIlla = 'Mallorca'
4 AND etapa.item = 'Grau superior';
```

```
1 SELECT nom
2 FROM centres centre t, t.nomEtapa etapa
3 WHERE nomIlla = 'Mallorca'
4 AND etapa.item = 'Grau superior';
```

Query History Saved Queries Results (11)

	nom
1	ACADEMIA FLEMING
2	ADEMA, ESCUELA DENTAL DE MALLORCA
3	ALBUHAIRA
4	ALCÚDIA
5	ANTONI MAURA
6	ARXIDUC LLUÍS SALVADOR
7	AULA BALEAR
8	BENDINAT
9	BERENGUER D'ANOIA
10	CALVIÀ
11	CALVIÀ

## Actualización 09/01/2025

Como se puede ver inicialmente realicé este ejercicio para el [segundo archivo JSON](#), el ya formateado para insertar los datos de forma directa en la tabla. En esta actualización del documento mostraré los cambios realizados en una repetición de la tarea (en otro servidor cloudera ya que el que se muestra en las capturas es uno que utilicé temporalmente en VirtualBox y no en Proxmox como suelo realizar y como es el caso en esta actualización) para los datos del [JSON original](#) de la tarea.

El primer cambio, evidentemente reside en la obtención del archivo. El comando `wget` se tiene que modificar la URL, resultando así en el siguiente (el resto de comandos son exactamente iguales<sup>1</sup>).

```
1 wget https://raw.githubusercontent.com/tnavarrete-iedib/bigdata-24-25/refs
2 /heads/main/centres_educatiu.json
```

La siguiente y última modificación con respecto al proceso anteriormente mostrado, es la propia inserción de los datos ya que se tiene primero que acceder al campo 'data' y extraer la información. Para ello, se ejecuta la siguiente sentencia en el editor de Hive:

```
1 INSERT INTO TABLE centres.centre
2 SELECT
3     get_json_object(centre, '$.adreca') AS adreca,
4     get_json_object(centre, '$.cif') AS cif,
5     get_json_object(centre, '$.codiIlla') AS codiIlla,
6     get_json_object(centre, '$.codiMunicipi') AS codiMunicipi,
7     get_json_object(centre, '$.codiMunicipi') AS codiMunicipi,
8     get_json_object(centre, '$.cp') AS cp,
9     CAST(get_json_object(centre, '$.esPublic') AS BOOLEAN) AS esPublic,
10    CAST(get_json_object(centre, '$.latitud') AS DOUBLE) AS latitud,
11    CAST(get_json_object(centre, '$.longitud') AS DOUBLE) AS longitud,
12    get_json_object(centre, '$.mail') AS mail,
13    get_json_object(centre, '$.nom') AS nom,
14    SPLIT(get_json_object(centre, '$.nomEtapa'), ',') AS nomEtapa,
15    get_json_object(centre, '$.nomIlla') AS nomIlla,
16    get_json_object(centre, '$.nomMunicipi') AS nomMunicipi,
17    get_json_object(centre, '$.telf1') AS telf1,
18    get_json_object(centre, '$.tipusCentreNomCa') AS tipusCentreNomCa,
19    get_json_object(centre, '$.web') AS web
20 FROM (
21     -- Provided in subject's forum
22     SELECT explode(
23         split(
24             regexp_replace(get_json_object(centre, '$.data'), '~\\[|\\]|$', ''),
25             '(?<=\\}|(?=\\{|)'
26         )
27     ) AS centre
28 FROM ext_centres
29 ) centres_exploded;
```

<sup>1</sup>Para mi caso en esta actualización recordar que era un servidor nuevo sin datos sobre esta actividad. En caso de haberlo sido, hubiera sido necesario borrar el primer archivo (`centres_educatiu2.json` o el nombre cambiado).



```

1 INSERT INTO TABLE centres.centre
2 SELECT
3   get_json_object(centre, '$.adreca') AS adreca,
4   get_json_object(centre, '$.cif') AS cif,
5   get_json_object(centre, '$.codiilla') AS codiilla,
6   get_json_object(centre, '$.codiMunicipi') AS codiMunicipi,
7   get_json_object(centre, '$.codiMunicipi') AS codiMunicipi,
8   get_json_object(centre, '$.cp') AS cp,
9   CAST(get_json_object(centre, '$.esPublic') AS BOOLEAN) AS esPublic,
10  CAST(get_json_object(centre, '$.latitud') AS DOUBLE) AS latitud,
11  CAST(get_json_object(centre, '$.longitud') AS DOUBLE) AS longitud,
12  get_json_object(centre, '$.mail') AS mail,
13  get_json_object(centre, '$.nom') AS nom,
14  SPLIT(get_json_object(centre, '$.nomEtopa'), ', ') AS nomEtopa,
15  get_json_object(centre, '$.nomilla') AS nomilla,
16  get_json_object(centre, '$.nomMunicipi') AS nomMunicipi,
17  get_json_object(centre, '$.telef') AS telef,
18  get_json_object(centre, '$.tipusCentreNomCa') AS tipusCentreNomCa,
19  get_json_object(centre, '$.web') AS web
20 FROM (
21   -- Provided in subject's forum
22   SELECT explode(
23     split(
24       regexp_replace(get_json_object(centre, '$.data'), '^\\[\\]\\$', ''),
25       '(?<=<\\))?(?=<\\D)'
26     )
27   ) AS centre
28 FROM ext_centres
29 ) centres_exploded;

```

	centre.adreca	centre.cif	centre.codiilla	centre.codir
1	JESÚS, 15	BS7134975	073	07040
2	Gremi Passamaners (Son Rossinyol), 11 2º	BS7178113	073	07040
3	BISBE BERENGUER DE PALOU, 6	S0718038C	073	07040
4	Gregori Méndel s/n (Parc Bit)	AS7358087	073	07040
5	MARIA I JOSEP, S/N	S0718151D	073	07039
6	ANTONI MARIA ALCOVER, S/N	S0718156C	073	07003
7	BENNASSAR, 8	S0718202E	073	07003

Como se puede observar en la imagen, el resultado es exactamente igual que en los pasos mostrados al comienzo de este documento para el archivo `centres_educatiu2.json`.