

Tarea evaluable

CE_5075 4.1

Big data aplicado



Índice

APARTADO 3 - Centros educativos de las Islas Baleares	2
1 IMPORTAR DATOS	3
2 CONSULTAS EN IMPALA	6
1. Recupera el número de centros públicos (esPublic será true) de la isla de Ibiza.	6
2. Recupera el nombre de todos los institutos de educación secundaria del municipio (nomMunicipi) de Palma.	7
3. Recupera el número de centros de cada tipo (tipusCentreNomCa) de la isla de Menorca.	7
4. Recupera el nombre de todos los centros de la isla de Mallorca que ofrecen estudios de la etapa (nomEtapa) "Grau superior".	8

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 lo puedes encontrar en el repositorio del curso. Descarga [este archivo](#) y súbelo a un directorio de HDFS.

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

Importante: No puedes editar el archivo previamente, debe cargarse tal y como está publicado.

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

Puedes trabajar con una copia del JSON correcto: [centres_educatius.json](#) (aunque solo contiene los primeros 100 centros).

1 IMPORTAR DATOS

Se han importado los datos de la URL proporcionada. En mi caso he usado directamente el archivo *centres_educatius.json* que contiene los datos dentro de “data”.

Después se ha creado la tabla en Hive además de la external table. Y por último se han insertado los datos extrayéndolos de “data”.

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

```
CREATE DATABASE centres;
```

```
CREATE TABLE centres.centre (  
  adreca STRING,  
  cif STRING,  
  codiIlla STRING,  
  codiMunicipi STRING,  
  codiOficial STRING,  
  cp STRING,  
  esPublic BOOLEAN,  
  latitud DOUBLE,  
  longitud DOUBLE,  
  mail STRING,  
  nom STRING,  
  nomEtapas ARRAY<STRING>,  
  nomIlla STRING,  
  nomMunicipi STRING,  
  telf1 STRING,  
  tipusCentreNomCa STRING,  
  web STRING  
) STORED AS PARQUET;
```

The screenshot shows the Hue web interface. On the left, a sidebar displays a file tree with 'centres' and 'centre' tables. The main area shows the SQL code for creating the 'centres.centre' table and the 'centres' database. Below the code, a 'Success' message is shown, followed by a 'Query History' section listing the executed queries with their timestamps and status.

```
1 CREATE TABLE centres.centre (  
2   adreca STRING,  
3   cif STRING,  
4   codiIlla STRING,  
5   codiMunicipi STRING,  
6   codiOficial STRING,  
7   cp STRING,  
8   esPublic BOOLEAN,  
9   latitud DOUBLE,  
10  longitud DOUBLE,  
11  mail STRING,  
12  nom STRING,  
13  nomEtapas ARRAY<STRING>,  
14  nomIlla STRING,  
15  nomMunicipi STRING,  
16  telf1 STRING,  
17  tipusCentreNomCa STRING,  
18  web STRING  
19 ) STORED AS PARQUET;
```

Success.

Query History

Time	Status	Query
a few seconds ago	✓	CREATE TABLE centres.centre (adreca STRING, cif STRING, codiIlla STRING, codiMunicipi STRING, codiOficial STRING, cp STRING, esPublic BOOLEAN, latitud DOUBLE, longitud DOUBLE, mail STRING, nom STRING, nomEtapas ARRAY<STRING>, nomIlla STRING, nomMunicipi STRING, telf1 STRING, tipusCentreNomCa STRING, web STRING) STORED AS PARQUET
4 minutes ago	✓	CREATE DATABASE centres

```
CREATE EXTERNAL TABLE ext_centres (centre STRING)
LOCATION '/user/cloudera/centres_educatius';
```



```
INSERT INTO TABLE centres.centre
SELECT
  get_json_object(centre, '$.adreca') AS adreca,
  get_json_object(centre, '$.cif') AS cif,
  get_json_object(centre, '$.codiIlla') AS codiIlla,
  get_json_object(centre, '$.codiMunicipi') AS codiMunicipi,
  get_json_object(centre, '$.codiMunicipi') AS codiMunicipi,
  get_json_object(centre, '$.cp') AS cp,
  CAST(get_json_object(centre, '$.esPublic') AS BOOLEAN) AS
esPublic,
  CAST(get_json_object(centre, '$.latitud') AS DOUBLE) AS
latitud,
  CAST(get_json_object(centre, '$.longituf') AS DOUBLE) AS
longitud,
  get_json_object(centre, '$.mail') AS mail,
  get_json_object(centre, '$.nom') AS nom,
  SPLIT(get_json_object(centre, '$.nomEtapa'), ', ') AS nomEtapa,
  get_json_object(centre, '$.nomIlla') AS nomIlla,
  get_json_object(centre, '$.nomMunicipi') AS nomMunicipi,
  get_json_object(centre, '$.telf1') AS telf1,
  get_json_object(centre, '$.tipusCentreNomCa') AS
tipusCentreNomCa,
  get_json_object(centre, '$.web') AS web
FROM (
  -- El siguiente código ha sido influido gracias al foro de
  preguntas de la asignatura
  SELECT explode(
    split(
      regexp_replace(get_json_object(centre, '$.data'),
        '^\\[[|\\]|$\\', ''),
        '(?<=\\}|, (?=\\{|)'
      )
    ) AS centre
  FROM ext_centres
) centres_split;
```

Add a name...
Add a description...

2m, 30s

centres

text

```

18 get_json_object(centre, '$.telef1') AS telef1,
19 get_json_object(centre, '$.tipusCentreNomCa') AS tipusCentreNomCa,
20 get_json_object(centre, '$.web') AS web
21 FROM (
22 -- El siguiente código ha sido influido gracias al foro de preguntas de la asignatura
23 SELECT explode(
24   split(
25     regexp_replace(get_json_object(centre, '$.data'), '^\\[\\]\\$', ''),
26     '{?<=\\}\\}\\{?(?=\\}\\{\\}')
27   )
28 ) AS centre
29 FROM ext_centres
30 ) centres_split;

```

Success.

Query History

Saved Queries

3 minutes ago

```

INSERT INTO TABLE centres.centre SELECT get_json_object(centre, '$.adreca') AS adreca,
get_json_object(centre, '$.cif') AS cif, get_json_object(centre, '$.codiIlla') AS codiIlla,
get_json_object(centre, '$.codiMunicipi') AS codiMunicipi,
get_json_object(centre, '$.codiMunicipi') AS codiMunicipi,
get_json_object(centre, '$.cp') AS cp,

```

Add a name...
Add a description...

0s

centres

text

```

1 SELECT * FROM centre;

```

Query History

Saved Queries

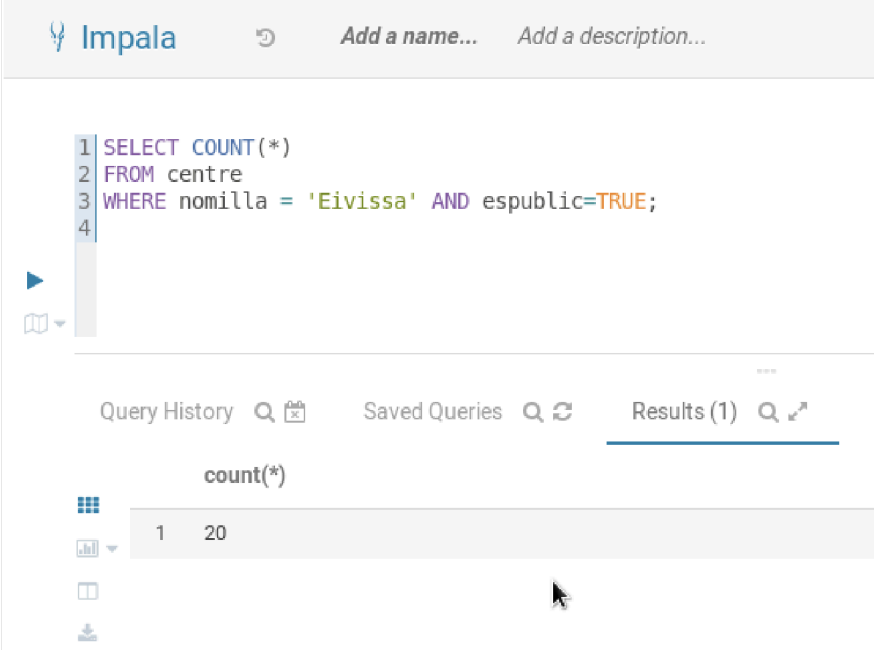
Results (100+)

	centre.adreca	centre.cif	centre.codiIlla	centre.codimunicipi	centre.codioficial
1	JESÚS, 15	B57134975	073	07040	07040
2	Gremi Passamaners (Son Rossinyol), 11 2º	B57178113	073	07040	07040
3	BISBE BERENGUER DE PALOU, 6	S0718038C	073	07040	07040
4	Gregori Méndel s/n (Parc Bit),	A57358087	073	07040	07040
5	MARIA I JOSEP, S/N	S0718151D	073	07039	07039
6	ANTONI MARIA ALCOVER, S/N	S0718156C	073	07003	07003

2 CONSULTAS EN IMPALA

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

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



The screenshot shows the Impala web interface. At the top, there's a header with the Impala logo and options to 'Add a name...' and 'Add a description...'. Below the header is a text area containing the SQL query:

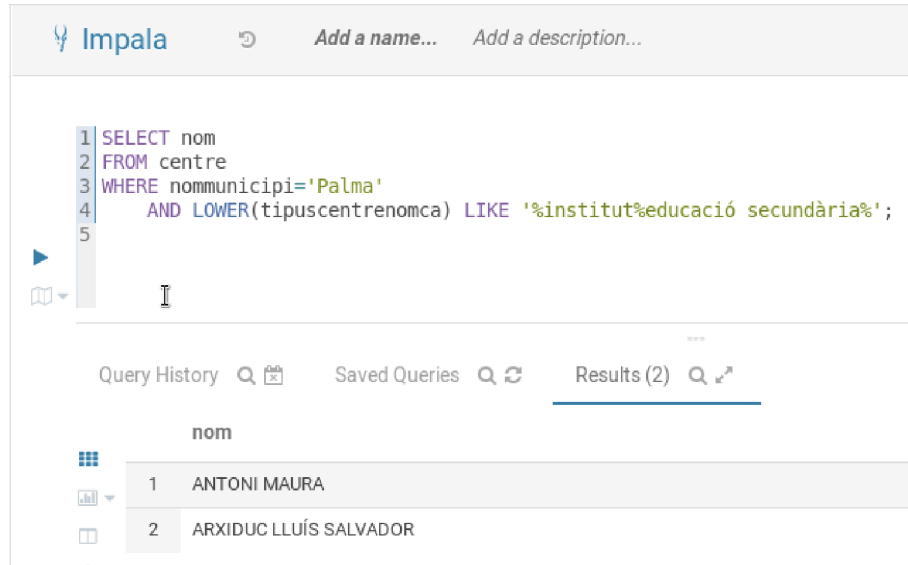
```
1 SELECT COUNT(*)  
2 FROM centre  
3 WHERE nomilla = 'Eivissa' AND espublic=TRUE;  
4
```

Below the query editor, there are tabs for 'Query History', 'Saved Queries', and 'Results (1)'. The 'Results (1)' tab is selected, showing a table with one column labeled 'count(*)' and one row with the value '20'.

count(*)
20

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

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



The screenshot shows the Impala query interface. The query is:

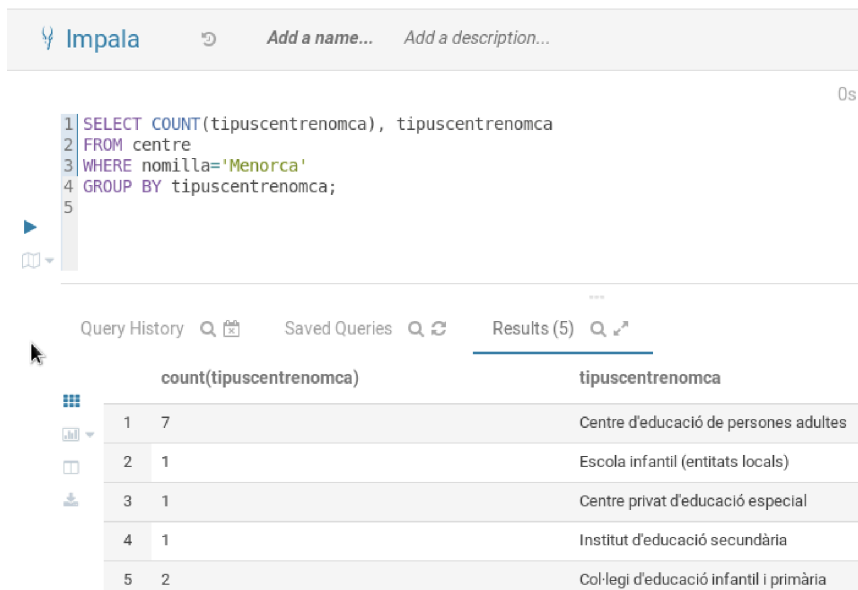
```
1 SELECT nom
2 FROM centre
3 WHERE nommunicipi='Palma'
4       AND LOWER(tipuscentrenomca) LIKE '%institut%educació secundària%';
5
```

The results are displayed in a table with the following data:

	nom
1	ANTONI MAURA
2	ARXIDUC LLUÍS SALVADOR

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

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



The screenshot shows the Impala query interface. The query is:

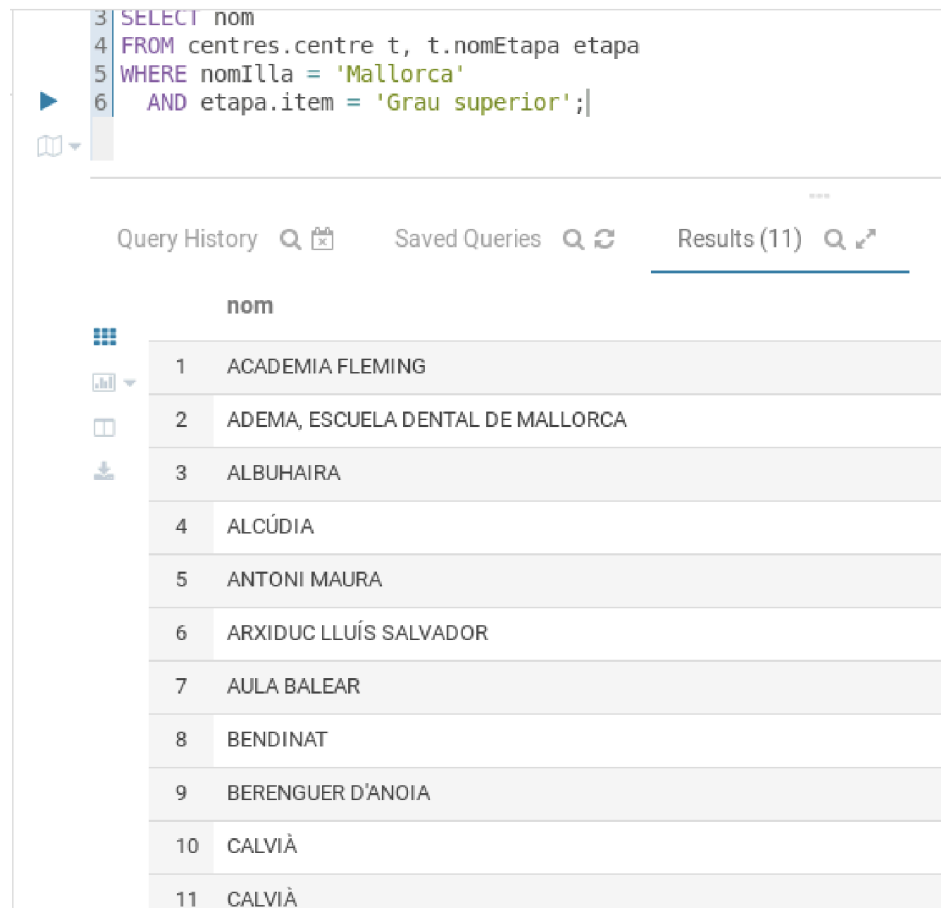
```
1 SELECT COUNT(tipuscentrenomca), tipuscentrenomca
2 FROM centre
3 WHERE nomilla='Menorca'
4 GROUP BY tipuscentrenomca;
5
```

The results are displayed in a table with the following data:

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

4. Recupera el nombre de todos los centros de la isla de Mallorca que ofrecen estudios de la etapa (nomEtapa) “Grau superior”.

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



The screenshot shows a database query interface. At the top, the SQL query is entered in a text area:

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

Below the query area, there are tabs for "Query History", "Saved Queries", and "Results (11)". The "Results (11)" tab is selected, showing a table with 11 rows. The table has a single column labeled "nom".

	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