

Informe de Pràctica sobre Spark SQL API - Anàlisi de Dades

Integrants:

- David Morillo Massagué (1666540)
 - Albert Guillaumet Mata (1672344)
 - Adrià Muro Gómez (1665191)
-

Introducció

En aquesta pràctica s'ha treballat amb l'eina Apache Spark, concretament amb la seva API SQL, per tal d'analitzar dades relacionades amb les receptes mèdiques facturades al Servei Català de la Salut ([CatSalut](#)).

Els conjunts de dades proporcionats, extrets del portal de dades obertes de la Generalitat de Catalunya, contenen informació sobre les prescripcions mèdiques gestionades per farmàcies, incloent dades com la regió sanitària, el sexe dels pacients, els medicaments prescrits, el nombre de receptes i el cost associat. Aquestes dades s'han treballat dins d'un entorn Google Colab, configurant un entorn Spark funcional, amb lectura i preparació dels fitxers *recetas.csv* i *header.csv*, per posteriorment dur a terme consultes analítiques mitjançant Spark SQL.

L'objectiu de la pràctica és aprendre a gestionar i analitzar grans volums de dades utilitzant Spark i la seva API SQL, aplicant consultes per extreure coneixement rellevant del dataset, com ara els medicaments més receptats, el cost total associat, les diferències entre sexes o les variacions segons la regió sanitària. A més, es resolen diverses preguntes analítiques amb consultes SQL per aprofundir en el coneixement del conjunt de dades.

Metodologia

Q1-How many drugs were prescribed during 2022, how many prescriptions and which was the overall cost?

```
consulta_q1 = """
SELECT
COUNT(DISTINCT medicament) AS num_medicaments,
SUM(CAST(nreceptes AS INT)) AS total_receptes,
SUM(CAST(import AS DOUBLE)) AS cost_total
FROM recetas2
WHERE any = 2022
"""
spark.sql(consulta_q1).show()
```

Explicació:

Es filtra per l'any 2022. Es compten medicaments únics, la suma total de receptes i l'import total. La conversió a DOUBLE del cost total assegura càlculs correctes del cost.

Resultat de la consulta:

1 row ▾ 1 rows x 3 cols		
num_medicaments	total_receptes	cost_total
367	23588623	3.1251608648000002E8

Q2-Which is the most prescribed drug in men and women and in which sanitary region?

```
consulta_q2 = """
SELECT sexe, rsanitaria, medicament, SUM(CAST(nreceptes AS INT)) AS total_receptes
FROM recetas2
GROUP BY sexe, rsanitaria, medicament
ORDER BY sexe, total_receptes DESC
"""
spark.sql(consulta_q2).show()
```

Explicació:

Fem groupBy per sexe, regió sanitària i medicament, sumem les receptes (*total_receptes*), ordenem per sexe i receptes en ordre descendent i filtrem la primera fila per sexe amb ROW_NUMBER().

Resultat de la consulta:

20 rows ▾ 20 rows × 4 cols

sexe	rsanitaria	medicament	total_receptes
Altres	ALTRES	AGONISTES OPIACIS	7546
Altres	ALTRES	Sense especificar	3506
Altres	ALTRES	Derivados de la b...	134
Altres	ALTRES	Inhibidores de la...	68
Altres	ALTRES	Inhibidores de la...	58
Altres	ALTRES	Anilidas	53
Altres	ALTRES	Inhibidores de la...	50
Altres	ALTRES	Inhibidores selec...	47
Altres	ALTRES	Vitamina D y anal...	45
Altres	ALTRES	Biguanidas	45
Altres	ALTRES	Agentes beta- blo...	41
Altres	ALTRES	Inhibidores de la...	37
Altres	ALTRES	Otros antihistami...	29
Altres	ALTRES	Derivados del aci...	27
Altres	ALTRES	Sulfonamidas, mon...	27
Altres	ALTRES	Otros agentes ant...	27
Altres	ALTRES	Tiazidas, monofar...	26
Altres	ALTRES	Pirazolonas	25
Altres	ALTRES	Penicilinas con e...	23
Altres	ALTRES	Adrenergicos en c...	20

Q3-Which is the least prescribed drug in men and women and in which sanitary region?

Consulta per dones:

```
consulta_q3d = """
SELECT sexe, rsanitaria, medicament, SUM(CAST(nreceptes AS INT)) AS total_receptes
FROM recetas2
WHERE sexe = 'Dona'
GROUP BY sexe, rsanitaria, medicament
ORDER BY total_receptes ASC
LIMIT 1
"""
spark.sql(consulta_q3d).show()
```

Consulta per homes:

```
consulta_q3h = """
SELECT sexe, rsanitaria, medicament, SUM(CAST(nreceptes AS INT)) AS total_receptes
```

```
FROM recetas2
WHERE sexe = 'Home'
GROUP BY sexe, rsanitaria, medicament
ORDER BY total_receptes ASC
LIMIT 1
"""
spark.sql(consulta_q3h).show()
```

Explicació: Utilitzem dues consultes, una per homes i un altre per dones. Les dues consultes filtren segons el sexe i s'agrupen amb el sexe, regió i medicament. Finalment, s'ordena a partir del total de receptes de manera ascendent i s'aplica *LIMIT 1* per veure el medicament menys receptat.

Resultat de la consulta:

1 row ▾	1 rows x 4 cols			
sexe	rsanitaria	medicament	total_receptes	
Dona	SENSE ESPECIFICAR	Derivados del tio...		1

1 row ▾	1 rows x 4 cols			
sexe	rsanitaria	medicament	total_receptes	
Home	TERRES DE L'EBRE	Derivados de la q...		1

Q4-Which is the most expensive drug prescribed?

```
consulta_q4 = """
SELECT medicament,
       SUM(CAST(import AS FLOAT)) AS total_import,
       SUM(CAST(nreceptes AS INT)) AS total_receptes,
       (SUM(CAST(import AS FLOAT)) / SUM(CAST(nreceptes AS INT))) AS
preu_mig_per_recepta
FROM recetas2
GROUP BY medicament
ORDER BY preu_mig_per_recepta DESC
LIMIT 1
"""
spark.sql(consulta_q4).show()
```

Explicació:

Per veure la droga més cara calculem el preu mitjà dividint l'import total amb el total de receptes. Agrupem per medicament i ordenem segons el preu mig de cada medicament en ordre descendent. Finalment limitem les sortides a 1 per veure el més car.

Resultat de la consulta:

medicament	total_import	total_receptes	preu_mig_per_recepta
Otras hormonas de...	1.1603210341308594E7	3213	3611.3321946182987

Q5-Considering the top 10 of most prescribed drugs during 2022, show the cost of the drugs for each sanitary region?

```
top10_query = """
SELECT medicament
FROM recetas2
WHERE any = 2022
GROUP BY medicament
ORDER BY SUM(CAST(nreceptes AS INT)) DESC
LIMIT 10
"""

spark.sql(top10_query).createOrReplaceTempView("top10_meds")

spark.sql("""
SELECT rsanitaria, medicament, SUM(CAST(import AS DOUBLE)) AS cost_total
FROM recetas2
WHERE any = 2022 AND medicament IN (SELECT medicament FROM top10_meds)
GROUP BY rsanitaria, medicament
ORDER BY rsanitaria, cost_total DESC
""").show(100)
```

Explicació:

Dividim el problema en dues consultes. La primera consulta filtra per l'any 2022, ordena els medicaments en ordre descendent segons el nombre de receptes i agafa els 10 amb més receptes. La segona consulta filtra per l'any 2022 i utilitza la consulta anterior per només agafar aquells medicaments del top 10 anterior. Aquests s'agrupen segons la regió i el medicament i s'ordenen segons el cost total en ordre descendent. Es limita la visualització a 100.

Resultat de la consulta:

88 rows ▾ 88 rows × 3 cols		
rsanitaria	medicament	cost_total
ALT PIRINEU i ARAN	Sense especificar	204690.41000000006
ALT PIRINEU i ARAN	Inhibidores de la...	76283.83000000002
ALT PIRINEU i ARAN	Inhibidores de la...	60614.150000000016
ALT PIRINEU i ARAN	Inhibidores selec...	41702.11
ALT PIRINEU i ARAN	Inhibidores de la...	33953.54000000001
ALT PIRINEU i ARAN	Derivados de la b...	29643.290000000005
ALT PIRINEU i ARAN	Anilidas	21570.919999999995
ALT PIRINEU i ARAN	Inhibidores de la...	19464.149999999998
ALT PIRINEU i ARAN	Derivados del aci...	17350.679999999997
ALT PIRINEU i ARAN	Agentes beta- blo...	13126.879999999996
ALTRES	Inhibidores de la...	56.55
ALTRES	Inhibidores selec...	31.35
ALTRES	Derivados de la b...	13.11
ALTRES	Inhibidores de la...	6.57
ALTRES	Inhibidores de la...	5.45
ALTRES	Agentes beta- blo...	4.9
ALTRES	Anilidas	2.5
ALTRES	Inhibidores de la...	1.45
BARCELONA	Sense especificar	1.7058155739999995E7
BARCELONA	Inhibidores de la...	6591390.040000001
BARCELONA	Inhibidores de la...	4486834.309999999
BARCELONA	Inhibidores selec...	3554964.220000001
BARCELONA	Inhibidores de la...	2257144.7899999996
BARCELONA	Anilidas	2244748.4400000004
BARCELONA	Derivados de la b...	1866789.3299999998
BARCELONA	Inhibidores de la...	1647302.0999999999
BARCELONA	Derivados del aci...	1206378.9300000004
BARCELONA	Agentes beta- blo...	1052633.0099999998

Conclusions

A través d'aquesta pràctica hem aprofundit en l'ús de l'API SQL de Spark com a eina principal per fer anàlisi de dades estructurades de manera eficient. El treball s'ha centrat en l'exploració un conjunt de dades reals relacionades amb les receptes mèdiques facturades pel Servei Català de la Salut, tot utilitzant consultes SQL sobre taules temporals creades amb DataFrames de Spark.

Els principals aprenentatges i conclusions han estat:

- Hem après a crear una sessió de Spark en un entorn col·laboratiu com Google Colab, i a preparar les dades perquè es puguin consultar fàcilment via SQL.
- Hem utilitzat instruccions SQL per filtrar, agrupar, ordenar i aplicar funcions agregades sobre columnes com el nombre de receptes i l'import econòmic.
- També hem après a fer càlculs derivats, com el cost mitjà per recepta, mitjançant conversions de tipus (CAST) i agregacions personalitzades.

En conjunt, la pràctica ens ha mostrat com Spark SQL pot oferir una manera clara i expressiva d'interactuar amb grans volums de dades sense necessitat d'utilitzar estructures imperatives. Això fa que sigui una eina molt potent per a l'anàlisi exploratòria i el reporting de dades en entorns Big Data.