Ampliant la caixa d'eines: common table expressions i funcions analítiques

i

El dimoni és en els detalls: optimització de la base de dades en funció del seu ús.

NOM I COGNOMS: _____

El Comitè Olímpic de la UOC, encarregat d'organitzar els *Jocs Atlètics Olímpics UOC*, ara que disposa de la base de dades que hem construït a la UOC com a part de l'assignatura **Bases de dades per a Data Warehousing**, ha obtingut el pressupost necessari per implementar una sèrie de millores. De nou, s'ha posat en contacte amb nosaltres perquè implementeu els requisits que ens han proposat.

Per a la proposta de solució d'aquesta PAC, heu de crear una base de dades nova anomenada **dbdw_pec4.** Primer executeu el script adjunt **DB_olympic_structure.sql**. A continuació excuteu el script **DB_olympic_data.sql.** Ambdós scripts SQL crearan l'estructura i el conjunt de dades necessaris pel desenvolupament d'aqueta PAC. NOTA: les dades proporcionades no es poden modificar i els exercicis han de fer-se amb aquestes dades.

Consideracions per al lliurament i realització de la PAC:

- Tot el que es demana en aquesta PAC està explicat en els blocs didàctics 4 i 5.
- Sigueu fidels al que diu l'enunciat. Si trobeu que algun enunciat no es prou clarificador o genera algun tipus de dubte, comenteu al fòrum o directament via correu al professor col·laborador.
- Es recomana la utilització de **pgAdmin** per a la implementació de tota la PAC. Hi ha una altra alternativa que és **psql** (línia de comandes), però és preferible que utilitzeu pgAdmin ja que és una interfície gràfica que us permetrà editar i crear sentències SQL (així com mostrar els resultats) de forma més senzilla que psql.
- Tal com s'indica a l'enunciat, cada resposta als exercicis ha d'entregar-se en un fitxer .sql diferent, amb el nom corresponent. S'avaluarà el codi lliurat en aquests fitxers .sql i **NO el codi que aparegui en el document amb les captures de pantalla**.
- Les captures de pantalla dels exercicis (i explicacions pertinents) han de proporcionar-se en un document a part (es proporciona una plantilla, indiqueu el vostre nom en el document).

- S'ha de realitzar el lliurament de tots els fitxers de la PAC (tant els fitxers .sql com el document amb explicacions i captures de pantalla) en un fitxer comprimit .zip.

Consideracions per a l'avaluació de l'exercici:

- Es tindrà en compte l'aplicació de les bones pràctiques de codificació en SQL, de consultes i de programació de procediments i disparadors, és a dir, codi amb sagnat, ús de clàusules SQL de forma correcta, comentaris, capçaleres en el procediment, etc.
- Els scripts proporcionats per l'estudiant amb les solucions dels exercicis han d'executar-se correctament. L'estudiant ha de assegurar-se que llançant l'script complet de cada exercici no es produeix cap error.
- Important: Les sentències SQL proporcionades en els scripts han de ser creades de forma manual i no mitjançant assistents que PostgreSQL/pgAdmin puguin proporcionar. Es pretén aprendre SQL i no la utilització d'assistents.
- Les sentències SQL proporcionades en els exercicis han de ser només aquelles que demana l'enunciat i cap altra més. Qualsevol sentència afegida addicionalment, si està malament o provoca que l'script no s'executi correctament a l'hora de corregir-lo, penalitzarà la puntuació de l'exercici.

EXERCICI 1 (30%)

El Comitè Olímpic de la UOC vol fer una sèrie de millores dins la base de dades. Ens han contactat per proposar els següents requeriments:

1) 10%) Es requereix que la taula *tb_athlete* tingui una columna *id* que s'ha d'anar inserint manualment amb un identificador que es va incrementant. El fet que es faci manualment pot generar errors. Per a resoldre-ho, es demana:

- Crear una seqüència *seq_athlete_id* que comenci per 1001 i que s'incrementi en 1
- Modificar la taula *tb_athlete* per afegir una columna *id* que faci servir per defecte la seqüència *seq_athlete_id* a la columna *id*.
- Actualitzar les dades existents (5403 registres). Per a actualitzar la columna id es farà servir una ordenació en base al camp *athlete_id* de forma descendent

2) (10%) Defineix, **mitjançant CTE recursives**, una consulta que retorni les posicions dels atletes a la competició de Triatló a la ronda 3. Han d'aparèixer la posició i el nom de l'atleta concatenar amb l'atleta que ha quedat a la posició anterior (ver imatge).

name character varying (50)	round_number integer	register_position integer	a_position text
Triathlon	3	0	0: ROYLE Aaron
Triathlon	3	1	0: ROYLE Aaron -> 1: SAGIV Ran
Triathlon	3	2	0: ROYLE Aaron -> 1: SAGIV Ran -> 2: CONINX Dorian
Triathlon	3	3	0: ROYLE Aaron -> 1: SAGIV Ran -> 2: CONINX Dorian -> 3: SANTOS Melanie
Triathlon	3	4	0: ROYLE Aaron -> 1: SAGIV Ran -> 2: CONINX Dorian -> 3: SANTOS Melanie -> 4: IDEN Gustav
Triathlon	3	5	0: ROYLE Aaron -> 1: SAGIV Ran -> 2: CONINX Dorian -> 3: SANTOS Melanie -> 4: IDEN Gustav -> 5: RAZARENOVA Alexandra
Triathlon	3	6	0: ROYLE Aaron -> 1: SAGIV Ran -> 2: CONINX Dorian -> 3: SANTOS Melanie -> 4: IDEN Gustav -> 5: RAZARENOVA Alexandra -> 6: PEREZ Irving
Triathlon	3	7	0: ROYLE Aaron -> 1: SAGIV Ran -> 2: CONINX Dorian -> 3: SANTOS Melanie -> 4: IDEN Gustav -> 5: RAZARENOVA Alexandra -> 6: PEREZ Irving -> 7: YEE Alex
Triathlon	3	8	0: ROYLE Aaron -> 1: SAGIV Ran -> 2: CONINX Dorian -> 3: SANTOS Melanie -> 4: IDEN Gustav -> 5: RAZARENOVA Alexandra -> 6: PEREZ Irving -> 7: YEE Ale…
Triathlon	3	9	0: ROYLE Aaron -> 1: SAGIV Ran -> 2: CONINX Dorian -> 3: SANTOS Melanie -> 4: IDEN Gustav -> 5: RAZARENOVA Alexandra -> 6: PEREZ Irving -> 7: YEE Ale…

3) (10%) Defineix, **mitjançant funcions analítiques**, la sentencia SQL que trobi tots els atletes identificant per cadascun d'ells:

- El nom de l'atleta, la disciplina i el país.
- Una columna que indiqui el millor temps de la disciplina. (REGISTER_TIME). Recordeu que algunes disciplines no tenen registre de temps per tant, haurien de tenir null al millor temps
- Una columna que indiqui el temps mig per país i disciplina. (REGISTER_TIME). Recordeu que algunes disciplines no tenen registre de temps per tant, haurien de tenir null al temps mig.
- Una columna que indiqui el nombre de participacions de l'atleta per país i disciplina. Un atleta pot participar més d'un cop i per tant, hauríeu de computar totes les participacions.
- Una columna que indiqui el nombre de participacions total per disciplina i país. Per tant, podem agrupar la informació de la columna anterior per país i disciplina.
- La informació s'ha d'ordenar per disciplina, país i atleta.

atleta character varying (50)	disciplina character varying (50)	pais character varying (3)	mejor_tiempo time without time zone	tiempo_medio_pais_disciplina interval	num_participaciones bigint	total_participantes_por_disciplina_pais numeric
PEROVA Ksenia	Archery	ROC	[null]	[null]	1	1
LEI Chien-Ying	Archery	TPE	[null]	[null]	1	7
TAN Ya-Ting	Archery	TPE	[null]	[null]	3	7
TANG Chih-Chun	Archery	TPE	[null]	[null]	2	7
WEI Chun-Heng	Archery	TPE	[null]	[null]	1	7
HAMMED Mohamed	Archery	TUN	[null]	[null]	1	1
MARCHENKO Veronika	Archery	UKR	[null]	[null]	1	1
ELLISON Brady	Archery	USA	[null]	[null]	1	2
KAUFHOLD Casey	Archery	USA	[null]	[null]	1	2
NGUYEN Hoang Phi Vu	Archery	VIE	[null]	[null]	1	1
HETHAT Yassine	Athletics	ALG	00:02:51	00:10:18	1	1
BISSET Catriona	Athletics	AUS	00:02:51	00:05:46.5	1	2
PASHLEY Ellie	Athletics	AUS	00:02:51	00:05:46.5	1	2
DEMIDIK Karyna	Athletics	BLR	00:02:51	00:10:42	1	1
SCOTCH Leungo	Athletics	BOT	00:02:51	00:08:54	1	1
BONFIM Caio	Athletics	BRA	00:02:51	00:12:14.25	1	4
CONSTANTINO Gabriel	Athletics	BRA	00:02:51	00:12:14.25	1	4

Les sentencies SQL dels 3 apartats s'han d'entregar en un fitxer amb nom **pec4_ej1.sql**.

La solució es pot trobar al fitxer pec4_ej1.sql

EXERCICI 2 (35%)

Donada la taula tb_register de la base de datos

athlete_id character (7)	round_number integer	discipline_id integer	register_date date	register_position integer	register_time time without time zone	register_measure real
1320573	0	1	2021-06-02	0	[null]	29.5
1304325	0	1	2021-06-02	1	[null]	29.57
1281372	0	1	2021-06-03	2	[null]	29.65
1378860	0	1	2021-06-04	3	[null]	29.75
1324035	0	1	2021-06-01	4	[null]	29.79
1429024	0	1	2021-06-02	5	[null]	29.82
1346893	0	1	2021-06-04	6	[null]	29.85
1324053	0	1	2021-06-03	7	[null]	29.89
1324193	0	1	2021-06-04	8	[null]	29.96
1323940	0	1	2021-06-01	9	[null]	30.02

Y donades les transaccions següents T1, T2, y T3:

T1:

```
START TRANSACTION;
  --S1.1:
  UPDATE olympic.tb_register SET register_measure = 29.75 WHERE athlete_id =
'1320573';
  --S1.2:
  SELECT register_measure FROM olympic.tb_register WHERE athlete_id = '1320573';
COMMIT;
```

T2:

```
START TRANSACTION;
  --S2.1:
  SELECT * FROM olympic.tb_register WHERE register_date > '2021-06-01';
  --S2.2:
  SELECT * FROM olympic.tb_register WHERE register_date > '2021-06-01';
COMMIT;
```

T3:

```
START TRANSACTION;
  --S3.1:
  SELECT * FROM  olympic.tb_register WHERE register_date = '2021-06-03' ;
  --S3.2:
  DELETE FROM olympic.tb_register WHERE athlete_id = '1281372';
  --S3.3:
  SELECT * FROM  olympic.tb_register WHERE register_date = '2021-06-03';
COMMIT;
```

a) (20%) Per a cada parella de transaccions, indiqueu les interferències que es poden produir i el nivell d'aïllament mínim per a evitar aquestes transferències. Justifica la teva resposta

T1 – T2: La transacció T1 no es veu afectada per la transacció T2, ja que aquesta última només fa SELECTs. La transacció T2 en canvi, sí que pot contenir una interferència de tipus 'lectura no repetible' produït per T1 (la interferència es produeix quan el athlete_id actualitza el valor del seu registre en S1.1 quan S2.1 ja havia realitzat la, i després del UPDATE en S1.2 la informació que apareix és diferent – register_measure). Per a evitar aquesta interferència, la transacció T2 hauria d'executar-se en manera REPEATABLE READ.

T1 – T3: La transacció T1 i T3 no s'interfereixen entre elles ja que accedeixen i manipulen línies de la taula diferents.

T2 – T3: La transacció T2 pot tenir una interferència de tipus 'lectura no repetible' causada per T3 (la interferència es produeix quan S3.2 elimina el athlete_id *1281372*, després que aquest hagi estat llegit en S2.1, i en realitzar novament la mateixa consulta en S2.2, aquest ja no apareix). Per a evitar aquesta interferència, la transacció T2 hauria d'executar-se'n manera REPEATABLE READ. La transacció T3 no es veu afectada per la transacció T2, ja que aquesta última només fa SELECTs.

b) (15%) Quin és el nivell mínim d'aïllament que hauria de tenir cada transacció per a garantir que no es produeixi cap tipus d'interferència amb qualsevol de les altres possibles transaccions.

T1: el seu nivell d'aïllament mínim hauria de ser SERIALIZABLE per a evitar que es puguin produir fantasmes. Per exemple, que una transacció actualitzi el register_date d'un atleta que abans era inferior al dia 2021-06-01 i després fos superior.

T2: el seu nivell d'aïllament mínim hauria de ser SERIALIZABLE per a evitar que es puguin produir fantasmes, com ja s'ha vist en l'apartat anterior.

T3: el seu nivell d'aïllament mínim hauria de ser SERIALIZABLE per a evitar que es puguin produir fantasmes. Per exemple, que una transacció afegeixi a un altre atleta amb la mateixa data de registre '2021-06-03'.

EXERCICI 3 (20%)

Des del Comitè Olímpic UOC ens demanen fer una tasca de recerca i pràctica sobre PostgreSQL el complement espacial PostGIS, com a experts que sou.

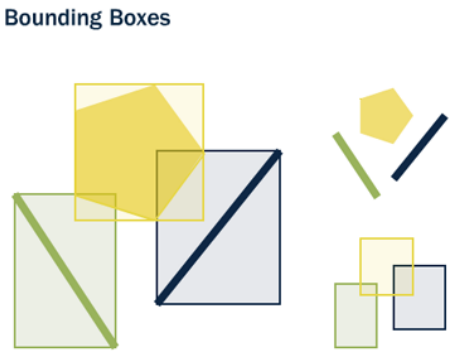
Volen conèixer més sobre els índex espacials per poder respondre les següents preguntes que s'estan fent:

- ¿Què son los índexs espacials?
- ¿Quan es fan servir?
- ¿Quins tipus en trobem?
- ¿Com es poden crear?

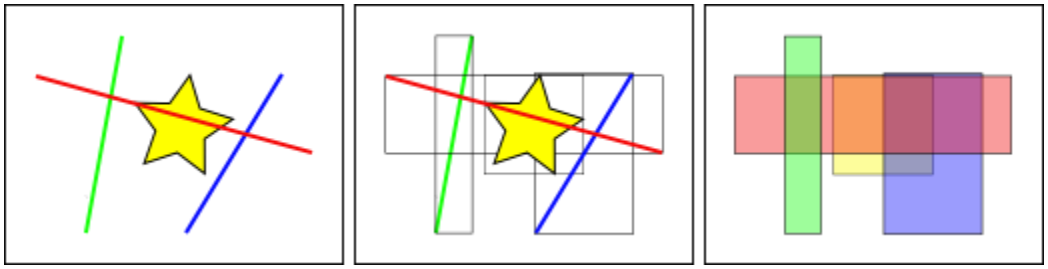
En cas necessari, l'explicació es podrà ampliar amb esquemes, imatges o exemples simples que ajudin a entendre més fàcilment la resposta proposada. ***(màxim 2 pàgines)***

La indexació espacial és una de les funcionalitats més importants de les bases de dades espacials. Els índexs aconsegueixen que les cerques espacials en un gran nombre de dades siguin eficients. Sense indexació, la cerca es realitzaria de manera seqüencial havent de buscar en tots els registres de la base de dades. La indexació organitza les dades en una estructura d'arbre que és recorreguda ràpidament en la cerca d'un registre.

Les base de dades estàndard creen un arbre jeràrquic basats en els valors de les columnes. Els índexs espacials funcionen d'una manera diferent, els índexs no són capaços d'indexar les geometries, i indexaran les caixes (box) de les geometries.



La caixa (box) és el rectangle definit per les màximes i mínimes coordenades X e I d'una geometria.



En la figura es pot observar que només la línia intersecta a l'estrella groga, mentre que si utilitzem els índexs comprovarem que la caixa groga és intersectada per dues figures la caixa vermella i la blava. El camí eficient per a respondre correctament a la pregunta quin element intersecta l'estrella groga? és primer respondre a la pregunta quines caixes intersecten la caixa groga? usant l'índex (consulta ràpida) i després calcular exactament qui intersecta a l'estrella groga? sobre el resultat de la consulta de les caixes.

Els índexs espacials s'utilitzen en PostgreSQL amb consultes que invoquen funcions per a comparar geometries. Les funcions ST_Geometry de PostgreSQL que usen índexs espacials són les que proven les relacions espacials.

- ST_Contains
- ST_Crosses
- ST_Disjoint
- ST_Equals
- ST_Intersects
- ST_Overlaps
- ST_Touches
- ST_Within

A més, la columna espacial ha d'aparèixer immediatament després de la funció de relació espacial en la consulta per a l'índex espacial que s'usarà.

EXERCICI 4 (15%)

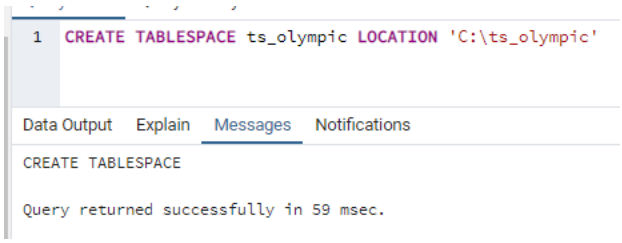
Donada la següent consulta SQL sobre la base de dades **dbdw_pec4**:

```
SELECT ta.name atleta,
          td.name disciplina,
          ta.country pais,
          tr.discipline_id,
          count(1)
FROM
    olympic.tb_register tr,
     olympic.tb_athlete ta,
     olympic.tb_discipline td
 WHERE
    1=1
    and ta.athlete_id = tr.athlete_id
    and td.discipline_id = tr.discipline_id
 GROUP by ta.name ,
          td.name ,
          ta.country ,
          tr.discipline_id
 ORDER BY ta.country, tr.discipline_id,ta.name
```

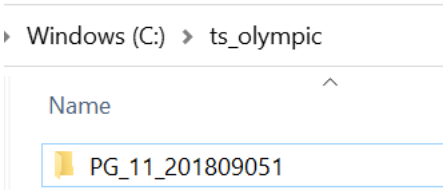
1. (5%) Creeu un espai de taules (*tablespace*) mitjançant una sentència SQL. Aquest tablespace ha d'anomenar-se ts_olympic i s'ha d'emmagatzemar a la ruta "x:\ts_olympic\", on x és la unitat de Windows on s'han d'emmagatzemar els fitxers del tablespace. Per exemple, x pot ser la unitat C:, la unitat F:, etc. (Nota: els que utilitzeu Linux podeu utilitzar la ruta /olympic/ en el directori arrel, o en Mac la ruta /Users/<el vostre usuari>/ts_olympic). Adjunteu una captura de pantalla del què PostgreSQL genera i expliqueu si heu hagut de realitzar alguna operació extra en el sistema operatiu.

   Nota: És possible que els que utilitzeu MAC i/o Linux hagueu d'assignar el permís "owner" a l'usuari "postgres". (Linux: sudo chown postgres /ts_clinical. Mac: sudo chown postgres /Users/<el vostre usuari/ts_clinical).

   S'ha de crear la carpeta ts_olympic, i després ja es pot executar la instrucción per a crear el tablespace



Es pot validar com al directori creat ha aparegut un nou fitxer



2. (5%) Identifica quin és el pla d'execució que fa servir PostgreSQL per resoldre aquesta consulta. Fes una captura de pantalla del pla i descriu-lo amb les teves paraules.

3. (5%) Creeu una vista materialitzada v participaciones a partir de la consulta especificada a aquest enunciat al tablespace ts_olympic i que generi les dades al crear-se

La solució dels apartats 2 i 3 es pot trobar al fitxer pec4_ej4.sql

Criteris de valoració

A l'enunciat s'indica el pes/valoració de cada exercici.

Per aconseguir la puntuació màxima en els exercicis, cal explicar amb claredat la solució que es proposa.

Format i data de lliurament

El format de l'arxiu que conté la vostra solució pot **ser .pdf, .doc i .docx**. **Per a altres opcions, si us plau, contacteu prèviament amb el vostre consultor.** El nom de l'arxiu ha de contenir el codi de l'assignatura, el vostre cognom i el vostre nom, així com el nombre d'activitat (PAC3).

El fitxer .zip que contingui tots els fitxers de la PAC (tant els fitxers .sql com el document que mostra els resultats de les vostres solucions) l'heu d'enviar a la bústia de Lliurament i registre d'AC disponible a l'aula (apartat Avaluació).

La data límit per lliurar la PAC4 és el **07/01/2022**.