# PRÁCTICA 3. NOSQL, OPTIMIZACIÓN Y TRANSACCIONES

Adrian San Felipe Martin – Luis Miguel Duran SISTEMAS INFORMÁTICOS 2020/2021

# Descripción del nuevo material entregado.

Además del código entregado en la práctica 2 de sistemas informáticos, se han añadido diferentes archivos y códigos:

- <u>createMongoDBFromPostgreSQLDB.py</u>: este archivo se encuentra dentro de la carpeta app, en la misma ruta que, por ejemplo, "routes.py" y es el script encargado de crear, a través de su ejecución con "python3 createMongoDBFromPostgreSQLDB.py" la base de datos si1 en mongoDB a partir de la base de datos proporcionada en postgres. Este script también crea la colección "topUSA" que contiene las 800 películas más actuales de Estados Unidos, ordenadas de más a menos actual. También contiene las funciones pedidas en el enunciado(buscar películas con "Life", películas que tengan como director a Woody Allen y películas que hayan participado juntos Parsons y Galecki)
- <u>Página html de topUsa</u>: nueva página web con los datos de las consultas mongoDB anteriores almacenadas en una tabla.
- <u>clientesDistintos.sql</u>: esta consulta sql muestra el número de clientes distintos que han realizado pedidos que superen un precio en un mes, por ejemplo, que número de clientes han realizado distintos pedidos en enero y que en esos mismos pedidos se hayan gastado más de 1000€.
- <u>countStatus.sql</u>: este script sql contiene las consultas, la creación de índices y las sentecias ANALYZE que se desarrollan en el apartado "2)Optimización" de la práctica.
- <u>Página html de borraCliente</u>: nueva página html que ejecuta una transacción que borra todos los datos de un cliente de la base de datos(incluido su carrito e historial de compra)
- updPromo.sql: script sql que crea una nueva columna promo en la tabla de customers.

# Resultados y respuestas de los ejercicios.

## 1)NoSQL

#### Apartado a)

En este apartado creamos la nueva base de datos a través de la proporcionada en postgres. La consulta para obtener las películas que nos piden es la siguiente (se encuentra en el archivo database.py):

En la primera consulta no recogemos los directores, ya que hay algunas películas que no tienen director y por tanto no mostraría esas mismas películas. Para solucionarlo, hemos creado otra consulta que mas adelante recoge los directores de cada película, la cual la pasamos por referencia.

#### Apartado b)

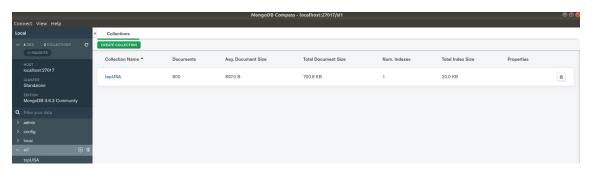
En este apartado se nos pide, con los datos anteriores, desarrollar el código para almacenar esos mismos datos en una base de datos en MongoDB llamada si1, que contenga una colección que se llame topUSA y con el código realizado, crear un script en Python llamado createMongoDBFromPostgreSQLDB.py

En la siguiente captura de pantalla se muestra el código desarrollado y su correcto funcionamiento:

```
import pymongo
import sys
import inspect
import os
current_dir = os.path.dirname(os.path.abspath(inspect.getfile(inspect.currentframe())))
parent_dir = os.path.dirname(current_dir)
sys.path.insert(0, parent_dir)
import database
client = pymongo.MongoClient("mongodb://localhost:27017/")
db = client["si1"]
if 'sil' in client.list_database_names():
   db.collection.drop()
   client.drop database("si1")
   print("La base de datos no estaba creada.")
collection = db["topUSA"]
movies = database.db_topUSA()
for item in movies:
   dict = {"title": item[0],
            year": item[1],
           "genres": item[2],
           "directors": "
           "actors": item[3]}
   directors=database.db getDirectors(item[0])
   if directors:
        for x in directors:
            dict["directors"]=dict["directors"]+str(x)
        dict["directors"]="Sin un director especifico"
   x = collection.insert_one(dict)
```

```
def mongo_Life():
    try:
       query=db.topUSA.find({"title":{"$regex":"Life"}});
        return list(query)
   except:
       print("Error Life")
        return []
def mongo_Allen():
       query=db.topUSA.find({"directors":{"$regex":"Allen, Woody"},"year":{"$regex":"199"}});
        return list(query)
   except:
       print("Error Allen")
       return []
def mongo_ParsonsGalecki():
       query=db.topUSA.find({"\actors":{\"\regex":"Parsons"}},{\"actors":{\"\regex":"Galecki"}}]});
        return list(query)
    except:
       print("Error Parsons")
        return []
```

Podemos ver en MongoDB Compass que ha creado la base de datos y la colección con 800 documentos.



## Apartado c)

En este apartado debemos crear un nuevo enlace en nuestra página web que conecte con la base de datos creada en MongoDB que se llame "Top USA", que mostramos a continuación:

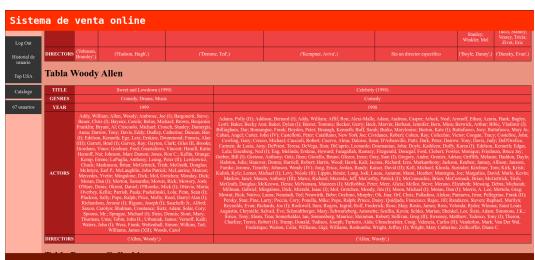


# Apartado d)

En este apartado tenemos que crear una nueva página html a la que nos llevará cuando pinchemos en el enlace anterior, y que nos muestre las 3 tablas pedidas en el enunciado, como se muestra a continuación:



Tu pagina web de confianza para comprar y alquilar DVDs - Web realizada por Luis Miguel y Adrian para SI



Tu pagina web de confianza para comprar y alquilar DVDs - Web realizada por Luis Miguel y Adrian para SI



## 2)Optimización

#### Apartado e)

En este apartado debemos estudiar el impacto de un índice mediante la creación de una consulta que se llama clientesDistintos.sql. Para realizar esta consulta se nos pedía que usásemos la sentencia EXPLAIN, que mejorásemos el rendimiento de la consulta, probar distintos índices y discutir los resultados. Después del proceso de discusión, los resultados son:

• Crear una consulta clientesDistintos.sql

```
select count(distinct customerid) as Suma
  from orders
  where date_part('month',orderdate)= 04 and
  date part('year',orderdate) = 2015 and totalamount > 100;
```

Sentencia EXPLAIN

Al hacer el explain, observamos que el coste de la consulta es de 5627.

Identificar un índice que mejore la consulta y crearlo

```
si1=# create index totalamount_index on orders (totalamount);
CREATE_INDEX
```

Hemos concluido que el índice que mejora el rendimiento es el que se realiza sobre totalamount, ya que al probar con orderdate el coste no variaba.

• Estudiar el nuevo plan de ejecución y compararlo con el anterior.

```
QUERY PLAN

Aggregate (cost=4480.32..4480.33 rows=1 width=8)
-> Bitmap Heap Scan on orders (cost=1126.90..4480.32 rows=2 width=4)
Recheck Cond: (totalamount > '100'::numeric)
Filter: ((date_part('month'::text, (orderdate)::timestamp without time z
one) = '4'::double precision) AND (date_part('year'::text, (orderdate)::timestamp
without time zone) = '2015'::double precision))
-> Bitmap Index Scan on totalamount_index (cost=0.00..1126.90 rows=605
97 width=0)
Index Cond: (totalamount > '100'::numeric)
(6 filas)
```

Como podemos observar, el coste de la consulta ha mejorado, pasando a valer 4480

Modificar el script clientesDistintos.sql

```
explain
select count(distinct customerid) as Suma
from orders
where date_part('month',orderdate)= 04 and
date part('year',orderdate) = 2015 and totalamount > 100;
```

## Apartado f)

Este apartado es únicamente de discusión y estudio, se trata principalmente de estudiar el impacto en la ejecución de unas sentencias sol que se nos incluyen en el guion de la práctica.

• Primera sentencia:

Segunda sentencia:

```
si1=# explain select customerid
si1-# from (
si1(# select customerid
si1(# from customers
si1(# union all
si1(# select customerid
si1(# from orders
si1(# where status='Paid'
si1(# ) as A
si1-# group by customerid
si1-# having count(*) =1;
                                 QUERY PLAN
HashAggregate (cost=4537.41..4539.41 rows=200 width=4)
   Group Key: customers.customerid
Filter: (count(*) = 1)
   -> Append (cost=0.00..4462.40 rows=15002 width=4)
         -> Seq Scan on customers (cost=0.00..493.93 rows=14093 width=4)
         -> Seq Scan on orders (cost=0.00..3959.38 rows=909 width=4)
               Filter: ((status)::text = 'Paid'::text)
(7 filas)
```

• Tercera sentencia:

Después de discutirlo, hemos decidimos crear un índice en orders(status) de la siguiente manera:

```
si1=# create index status_orders on orders(status);
CREATE INDEX
```

Y volvemos a lanzar las consultas de nuevo después de crear el índice.

Primera consulta:

Segunda consulta:

```
si1=# explain select customerid
si1-# from (
si1(# select customers
si1(# from customers
si1(# union all
si1(# select customerid
si1(# from orders
si1(# where status='Paid'
si1(#) as A
si1-# group by customerid
si1-# having count(*) =1;

QUERY PLAN

HashAggregate (cost=2074.55..2076.55 rows=200 width=4)
Group Key: customers.customerid
Filter: (count(*) = 1)
-> Append (cost=0.00..1999.54 rows=15002 width=4)
-> Seq Scan on customers (cost=0.00..493.93 rows=14093 width=4)
-> Bitmap Heap Scan on orders (cost=19.46..1496.52 rows=909 width=4)
Recheck Cond: ((status)::text = 'Paid'::text)
-> Bitmap Index Scan on status_orders (cost=0.00..19.24 rows=909 width=0)
Index Cond: ((status)::text = 'Paid'::text)
(9 filas)
```

Tercera consulta:

Como podemos observar, hemos logrado una mejoría en los tiempos a la hora de añadir el índice anterior.

- ¿Qué consulta devuelve algún resultado nada más comenzar su ejecución?
   La última consulta, ya que su coste al comenzar es prácticamente de cero.
- ¿Qué consulta se puede beneficiar de la ejecución en paralelo?
   En la segunda consulta, ya que las subconsultas que contiene se pueden paralelizar.

#### Apartado g)

Este apartado se basa en estudio del impacto de la generación de estadísticas, discusión y creación de un script countStatus.sql. Para comenzar, debemos estudiar y discutir las sentencias que se nos incluyen en el guion de la práctica (antes de ejecutar las sentencias hemos reiniciado la base de datos):

• Primera consulta:

• Segunda consulta:

Tras crear el índice en orders(status):

```
si1=# create index status_orders on orders(status);
CREATE INDEX
```

Primera consulta:

```
si1=# explain select count(*)
si1-# from orders
si1-# where status is null;

QUERY PLAN

Aggregate (cost=1496.52..1496.53 rows=1 width=8)
-> Bitmap Heap Scan on orders (cost=19.46..1494.25 rows=909 width=0)
Recheck Cond: (status IS NULL)
-> Bitmap Index Scan on status_orders (cost=0.00..19.24 rows=909 width=0)
Index Cond: (status IS NULL)
(5 filas)
```

Segunda consulta:

Generamos las estadísticas:

## si1=# analyze; ANALYZE

ANALYZE sin ningún parámetro analiza todas las tablas de la base de datos actual. Ahora, ejecutamos las dos consultas anteriores de nuevo más las dos siguientes que se nos indican en el guion.

• Primera consulta:

```
si1=# explain select count(*)
si1-# from orders
si1-# where status is null;

QUERY PLAN

Aggregate (cost=7.26..7.27 rows=1 width=8)
-> Index Only Scan using status_orders on orders (cost=0.42..7.26 rows=1 width=0)
Index Cond: (status IS NULL)
(3 filas)
```

Segunda consulta:

• Tercera consulta:

Cuarta consulta:

Conclusiones: El generador de estadísticas "ANALYZE" sirve para analizar los datos de la base de datos o tabla que indiquemos para optimizar las búsquedas. La diferencia entre las consultas antes de hacer el analyze y después viene dada porque tras analizar los datos, cuando hay un status que ocupa un gran porcentaje de las filas, es ineficiente tenerlo en el índice, ya que es más costoso mantenerlo que buscar en todas las filas.

## 3) Transacciones

#### Apartado h)

Como ejemplo de transacciones en la base de datos, queremos borrar los datos de usuario con ID 15. Para empezar, provocaremos un fallo de integridad que anule la transacción y haga rollback.

## Ejemplo de Transacción con Flask SQLAlchemy

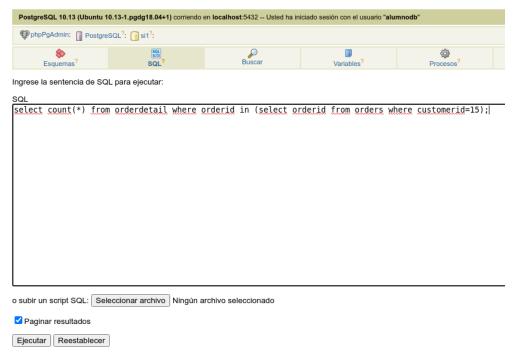
Customer ID: 15
<ul> <li>Transacción vía sentencias SQL</li> <li>Transacción vía funciones SQLAlchemy</li> </ul>
☐ Ejecutar commit intermedio
✓ Provocar error de integridad
Duerme 0 segundos (para forzar deadlock).
Enviar
Trazas

Al darle a enviar, nos saldrán estas trazas:

#### **Trazas**

- 1. Se ejecuta a traves de SQL
- 2. BEĞIN
- 3. Se borrarán los datos de los pedidos del cliente: 15
- 4. Error en la transaccion. Haciendo rollback.
- Cerramos la conexión con la base de datos.

Podemos comprobar que los datos de los pedidos efectivamente no se han borrado a través de phpPGAdmin:



Y como podemos observar, siguen saliendo los pedidos:



Ahora vamos a probar a ejecutarlo con un fallo, pero haciendo commit intermedio:

## Ejemplo de Transacción con Flask SQLAlchemy

Customo	er ID:	1					
			encias S( ciones SQ	•	emy		
□Ejecu	tar coı	mmit inte	ermedio				
✓ Prove	ocar en	ror de int	egridad				
Duerme	0			segun	dos (para	a forzar o	deadlock)
Enviar							

### Trazas

- 1. Se ejecuta a traves de SQL
- 2. BEGIN
- 3. Se borrarán los datos de los pedidos del cliente: 15
- 4. Commit intermedio. Se inicia otra transaccion
- 5. Error en la transaccion. Haciendo rollback.
- 6. Cerramos la conexión con la base de datos.

Ahora comprobamos si se han borrado los pedidos en la base de datos con la sentencia anterior:



Al hacer el commit intermedio, el rollback vuelve al momento después de borrar los detalles de los pedidos.

Vamos a probar a borrar otro usuario con un id distinto y sin provocar ningún error:

## Ejemplo de Transacción con Flask SQLAlchemy

Como podemos observar, la transacción se ha realizado correctamente.

Si intentamos borrar de nuevo el mismo cliente, hará un rollback, ya que ya no existe.

## Ejemplo de Transacción con Flask SQLAlchemy

Customer ID: 7
<ul> <li>Transacción vía sentencias SQL</li> <li>Transacción vía funciones SQLAlchemy</li> </ul>
☐ Ejecutar commit intermedio
✓ Provocar error de integridad
Duerme o segundos (para forzar deadlock).
Enviar
Trazas
Se ejecuta a traves de SQL     BEGIN     Frror en la transaccion, Haciendo rollback

4. Cerramos la conexión con la base de datos.

(El funcionamiento con SQLAlchemy es el mismo que con SQL)

## Apartado i)

Este es el script de creación del trigger desarrollado:

```
--Apartado b) creacion de la nueva columna promo
alter table customers add promo integer;
--Apartado c) creacion del trigger
CREATE OR REPLACE FUNCTION promo() returns TRIGGER
language plpgsql
as $$
DECLARE
BEGIN
update orders set totalamount = totalamount - totalamount * promo/100
    inner join orders on customer.customerid=orders.customerid
   where customerid= NEW.customerid and status = null;
    return new;
    --Apartado d) modificar el trigger para que haga un sleep con pg_sleep
   perform pg sleep(15);
END;
$$;
create trigger updPromo after update or insert on customers
for each row execute procedure promo();
```

Como podemos observar, se crea correctamente:

```
si1=# \i app/sql/updPromo.sql
ALTER TABLE
CREATE FUNCTION
CREATE TRIGGER
```

## • Explicación del deadlocks y como se pueden evitar:

El deadlock se produce al realizar una transacción de borrado, ya que intenta borrar los datos de la tabla customers, pero esta tabla está bloqueada por el trigger que hemos creado en updPromo.sql, ya que realiza una actualización.

Podemos evitar los deadlocks reproduciendo el tiempo de ejecución de las transacciones y optimizarlas con índices.