

**Máster universitario de Ciencia de Datos**

**Práctica 1 – PRA1**

**Uso de base de datos NoSQL - Cassandra, MongoDB y Neo4j.**

Autor:

Mario Ubierna San Mamés

|  |
| --- |
| Índice de Contenido |

[Índice de Contenido 3](#_Toc89849666)

[Índice de tablas 4](#_Toc89849667)

[Índice de ilustraciones 5](#_Toc89849668)

[1. Ejercicio 1: Cassandra 6](#_Toc89849669)

[2. Ejercicio 2: MongoDB 7](#_Toc89849670)

[3. Ejercicio 3: Neo4j 13](#_Toc89849671)

[4. Ejercicio 4: Neo4j 19](#_Toc89849672)

[5. Bibliografía 22](#_Toc89849673)

|  |
| --- |
| Índice de tablas |

**No se encuentran elementos de tabla de ilustraciones.**

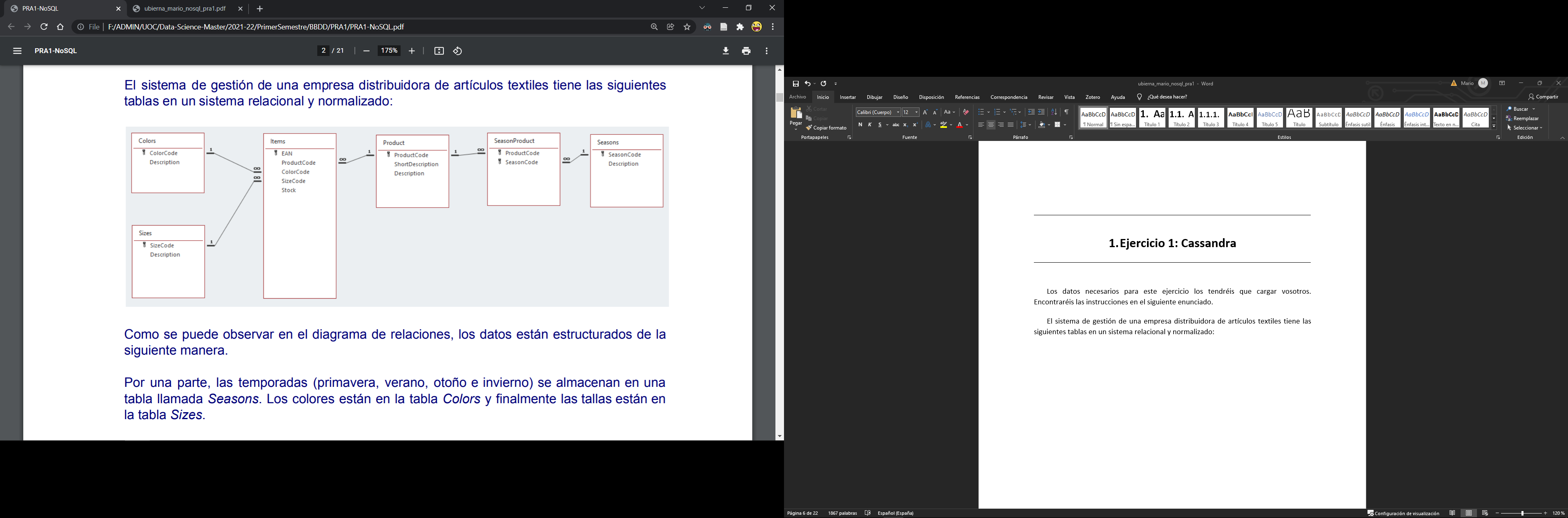
|  |
| --- |
| Índice de ilustraciones |

**No se encuentran elementos de tabla de ilustraciones.**

|  |
| --- |
| Ejercicio 1: Cassandra |

*Los datos necesarios para este ejercicio los tendréis que cargar vosotros. Encontraréis las instrucciones en el siguiente enunciado.*

*El sistema de gestión de una empresa distribuidora de artículos textiles tiene las siguientes tablas en un sistema relacional y normalizado:*



*Como se puede observar en el diagrama de relaciones, los datos están estructurados de la siguiente manera.*

*Por una parte, las temporadas (primavera, verano, otoño e invierno) se almacenan en una tabla llamada Seasons. Los colores están en la tabla Colors y finalmente las tallas están en la tabla Sizes.*

*La tabla Product contiene el código de los productos, una descripción corta para los tickets de compra y albaranes, y una descripción más detallada para la web comercial. Un producto puede estar disponible en diferentes tallas y colores.*

*Como un producto puede pertenecer a más de una temporada, hay una tabla intermedia llamada SeasonProduct que relaciona el producto con todas las temporadas a las que pertenece. Es una relación de muchos a muchos entre temporadas y productos.*

*En la tabla Item es donde se establecen las unidades disponibles de cada producto. Cada fila de la tabla representa una composición de un producto (ProductCode), una talla (SizeCode) y un color (ColorCode). Por reglas de integridad, esta composición es única, no puede repetirse. Al ser el nivel más bajo de detalle, cuenta con el stock disponible y una referencia única llamada EAN que es la clave primaria de la tabla Item. Por lo tanto, pueden haber diferentes colores y tallas de un mismo producto, es decir el campo ProductCode puede repetirse en la tabla Item. Sin embargo, un EAN no puede repetirse en la tabla Item ya que es clave primaria.*

## Ejercicio 1.1 (no puntúa)

Mirar el enunciado de la práctica (carga de los datos).

## Ejercicio 1.2 (80%)

*Con los datos que habéis cargado en el apartado anterior, se requiere realizar las siguientes consultas y adjuntar su resultado. Limitaciones para las consultas:*

* *Para cada consulta deberéis elegir la tabla más adecuada.*
* *No se pueden buscar manualmente en la base de datos los códigos necesarios para las consultas de actualización. Se deben recuperar los datos solicitados a partir de la información facilitada en el enunciado. En un escenario real con muchos registros no es viable seleccionar toda una tabla entera y buscar los datos manualmente en pantalla.*
* *No se puede utilizar la cláusula ALLOW FILTERING.*

### Consulta 1 (10%)

*El departamento de logística quiere saber el stock total de todos los productos que hay en el almacén. La consulta debe retornar una sola cifra. El ejercicio devuelve un warning. ¿Podrías explicar por qué?*

La consulta es:

*select sum(stock) from product\_color\_size\_stock ;*

El resultado de la consulta:

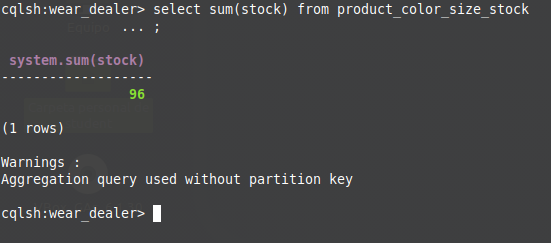


Ilustración 1 - Cassandra consulta y resultado consulta 1.

En este caso al ejecutar la consulta se produce un *warning*, esto se debe a que estamos haciendo uso de una *select* en la que el agregado es definido por nosotros sin una clave de partición, es decir, estamos haciendo una suma de todos los valores que hay en *stock*. Esto puede ser contraproducente si hay muchas filas, ya que para resolver esta consulta se tiene que leer todas las filas que hay en la tabla, pudiendo llegar al tiempo máximo establecido para una consulta y afectar al rendimiento.

Por lo tanto, lo que nos dice este *warning* es que es mejor hacer uso de una clave de partición, para así solo leer los datos de la clave de partición que se indique, mejorando el rendimiento de la consulta y del servidor de *Cassandra*.

### Consulta 2 (15%)

*El mismo departamento quiere saber el stock del artículo con referencia ‘CASTR’ agrupado por colores. La consulta debe detallar el código del producto, el código del color y el stock.*

La consulta realizada es:

*SELECT productcode, colorcode, SUM(stock)*

*FROM product\_color\_size\_stock*

*WHERE productcode = 'CASTR'*

*GROUP BY colorcode*

*;*

El resultado de la consulta es:

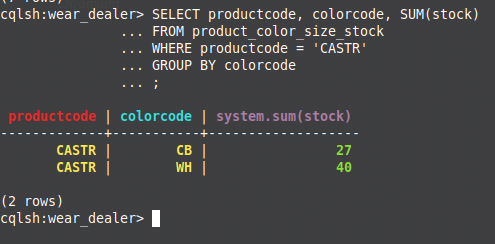


Ilustración 2 - Cassandra consulta y resultado consulta 2.

### Consulta 3 (20%)

*Listar los productos sin stock para hacer nuevos pedidos a fábrica. Se requieren el código del producto, el código del color, el código de talla y el EAN.*

La consulta es:

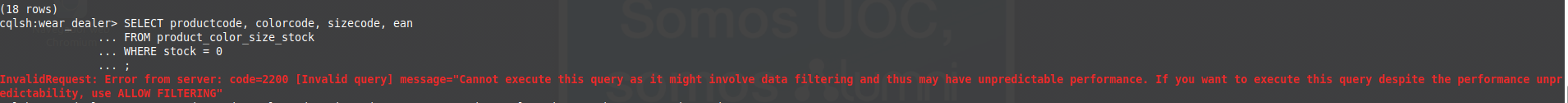
*SELECT productcode, colorcode, sizecode, ean*

*FROM product\_color\_size\_stock*

*WHERE stock = 0*

*;*

El resultado que nos genera:



Como podemos apreciar nos devuelve un error del tipo *code=2200 Invalid query*. Para solventar este problema hemos realizado lo siguiente, crear un índice para esa columna y volver a ejecutar la consulta de antes. Creamos el índice:

*CREATE INDEX iStock on product\_color\_size\_stock(stock) ;*

Ejecutamos la consulta anterior de nuevo, proporcionando el siguiente resultado:

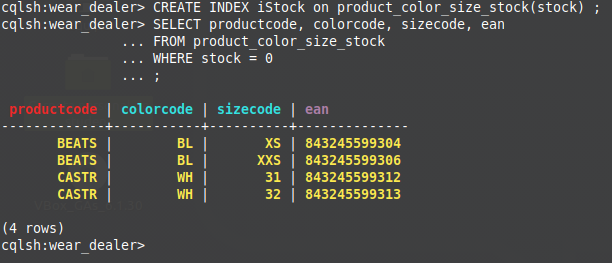


Ilustración 3 - Cassandra consulta y resultado consulta 3.

### Consulta 4 (15%)

*El departamento de marketing quiere revisar el catálogo de productos. Necesitan responder a la pregunta ¿Cuáles son los productos que hay en cada temporada? Se necesita un listado que proporcione Código de temporada, código de producto y la descripción corta. No necesitan ningún recuento de stock.*

La consulta realizada es:

*SELECT seasoncode, productcode, shortdescription*

*FROM seasons\_product\_color\_size*

*GROUP BY seasoncode, productcode*

*;*

El resultado que nos genera es el siguiente:

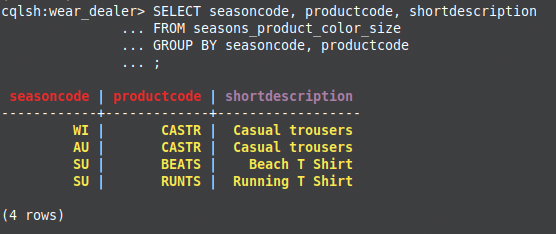


Ilustración - Cassandra consulta y resultado consulta 4.

### Consulta 5 (40%)

*Actualizar la descripción del producto con código referencia RUNTS para todas las tallas de manera consistente para toda la base de datos. La nueva descripción será ‘Running T-Shirt High Performance’.*

*La consulta realizada para la tabla product\_color\_size\_stock es:*

*SELECT productcode, colorcode, sizecode, shortdescription*

*FROM product\_color\_size\_stock*

*WHERE productcode = 'RUNTS'*

*;*

*UPDATE product\_color\_size\_stock*

*SET shortdescription = 'Running T-Shirt High Performance'*

*WHERE productcode = 'RUNTS' and colorcode = 'YE' and sizecode = 'S'*

*;*

*UPDATE product\_color\_size\_stock*

*SET shortdescription = 'Running T-Shirt High Performance'*

*WHERE productcode = 'RUNTS' and colorcode = 'YE' and sizecode = 'M'*

*;*

*UPDATE product\_color\_size\_stock*

*SET shortdescription = 'Running T-Shirt High Performance'*

*WHERE productcode = 'RUNTS' and colorcode = 'YE' and sizecode = 'L'*

*;*

*UPDATE product\_color\_size\_stock*

*SET shortdescription = 'Running T-Shirt High Performance'*

*WHERE productcode = 'RUNTS' and colorcode = 'YE' and sizecode = 'XL'*

*;*

La cual genera como resultado:

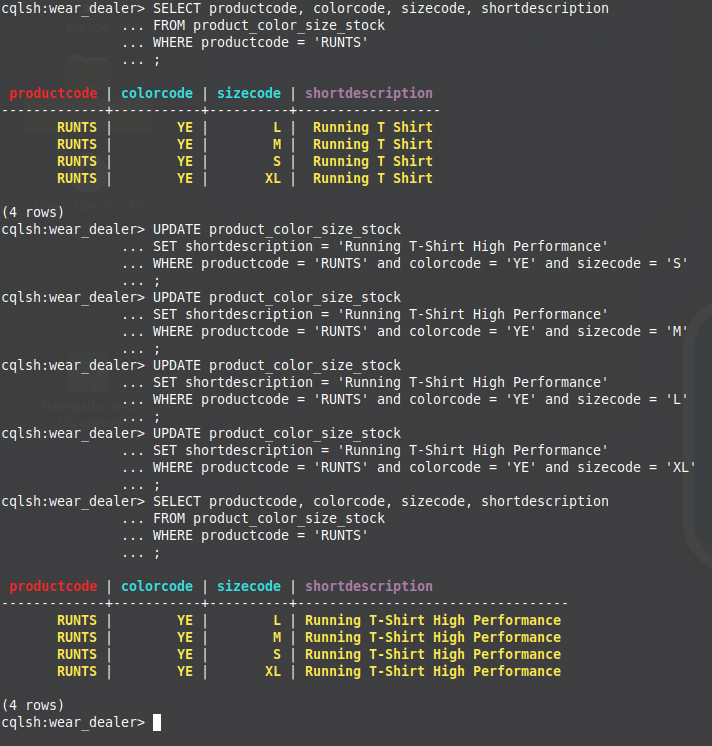


Ilustración - Cassandra consulta y resultado consulta 5.

|  |
| --- |
| Ejercicio 2: MongoDB |

*Para este ejercicio considera la base de datos descrita en el documento “Diseño de una base de datos para una app de mensajería instantánea” que se encuentra en los materiales del curso.*

*También se necesitará la máquina virtual LinuxMint que contiene una instalación de MongoDB y la base de datos ya cargada. Para este ejercicio no es necesario cargar ningún dato adicional.*

*Se pide proporcionar las sentencias (en texto) para el shell de MongoDB y los resultados que se obtienen (haciendo una captura de pantalla o adjuntando el texto retornado) para las siguientes consultas:*

## Consulta 1 (20%)

*De la colección Contactos listar los documentos que tengan un contacto cuya edad sea igual o mayor que 52 años ordenado por identificador (no el campo “\_id”) ascendente. El listado debe contener el identificador (no el campo “\_id”), el nombre y los apellidos del usuario, y el nombre y los apellidos del contacto.*

La consulta es:

*db.Contactos.find(*

*{*

*"Contacto.Edad": {$gte: 52}*

*},*

*{*

*\_id: 0,*

*Identificador: 1,*

*"Usuario.Nombre": 1,*

*"Usuario.Apellidos": 1,*

*"Contacto.Nombre": 1,*

*"Contacto.Apellidos": 1,*

*}*

*).sort(*

*{*

*Identificador: 1*

*}*

*)*

Y el resultado de la consulta:

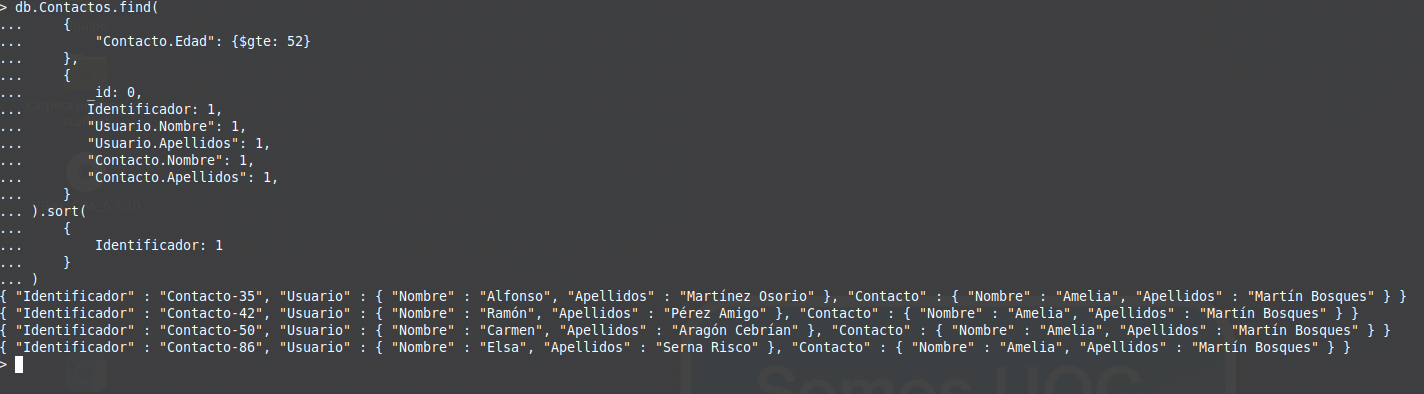


Ilustración 6 - MongoDB consulta y resultado de la consulta 1.

## Consulta 2 (20%)

*De la colección Desbloqueos listar los desbloqueos realizados por el usuario con email "jsanzrobles@hotmail.es". El listado debe mostrar el nombre y el email del usuario desbloqueado, y el identificador del desbloqueo (no el campo “\_id”).*

La consulta es:

*db.Desbloqueos.find(*

*{*

*"Usuario\_desbloqueador.Email": "jsanzrobles@hotmail.es"*

*},*

*{*

*\_id: 0,*

*Identificador: 1,*

*"Usuario\_desbloqueado.Nombre": 1,*

*"Usuario\_desbloqueado.Email": 1*

*}*

*)*

El resultado de la consulta es:

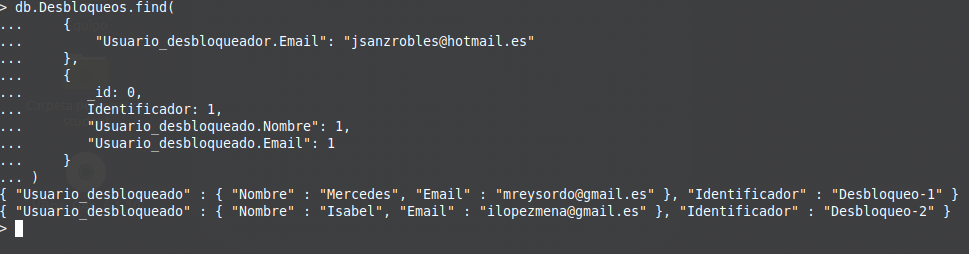


Ilustración 7 - MongoDB consulta y resultado de la consulta 2.

## Consulta 3 (30%)

*De la colección Usuarios\_grupos listar los tres usuarios que son propietarios de más grupos en orden descendente (el usuario que tiene más grupos en propiedad debe aparecer el primero). La consulta debe retornar solamente los tres primeros, mostrar el correo electrónico del proprietario y el recuento de los grupos de los que es propietario. Se puede asumir que no hay correos electrónicos repetidos en esta colección, no es necesario realizar ninguna comprobación adicional.*

La consulta es:

*db.Usuarios\_grupos.aggregate([*

*{*

*$project: {*

*\_id: 0,*

*Email: 1,*

*Count: {$size: "$Grupos\_propietario"}*

*}*

*},*

*{*

*$sort: {"Count": -1}*

*},*

*{*

*$limit: 3*

*}*

*])*

El resultado de la consulta es:

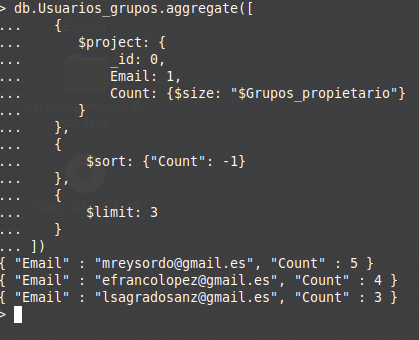


Ilustración 8 - MongoDB consulta y resultado de la consulta 3.

## Consulta 4 (30%)

*De la colección Contactos\_usuarios y del usuario con email mgarciasanz@gmail.es, seleccionar solamente los contactos de este usuario que tengan una edad igual o mayor que 36 años. La consulta debe retornar el email del usuario (mgarciasanz@gmail.es), y el email y edad de sus contactos. Los contactos de este usuario que no estén en el rango de edad especificado no deben salir en la consulta.*

La consulta hecha es:

*db.Contactos\_usuarios.aggregate([*

*{*

*$match: {"Email": "mgarciasanz@gmail.es"}*

*},*

*{*

*$unwind: "$Contactos"*

*},*

*{*

*$match: {"Contactos.Usuario\_contacto.Edad": {$gte: 36}}*

*},*

*{*

*$project: {*

*\_id: 0,*

*Email: 1,*

*"Contactos.Usuario\_contacto.Email": 1,*

*"Contactos.Usuario\_contacto.Edad": 1*

*}*

*}*

*])*

El resultado que proporciona:

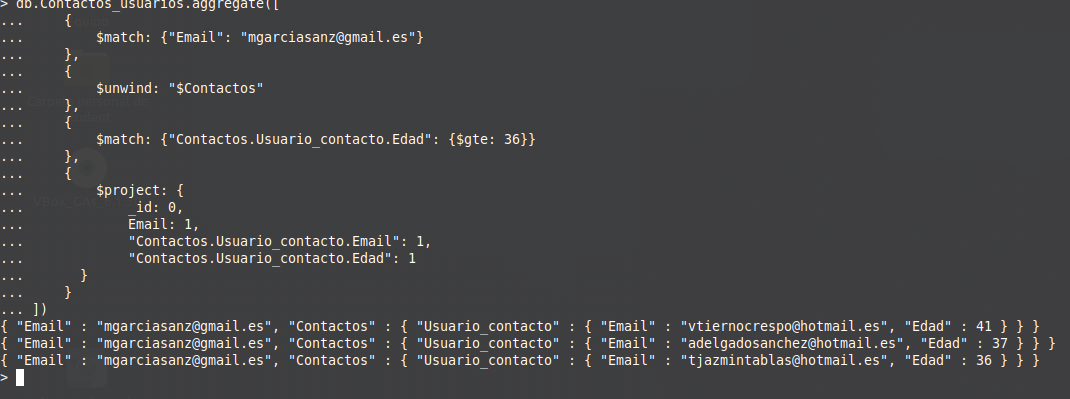


Ilustración 9 - MongoDB consulta y resultado de la consulta 4.

|  |
| --- |
| Ejercicio 3: Neo4j |

*Para la realización de este ejercicio se seguirán las instrucciones del caso de estudio ubicado en la siguiente URL: https://neo4j.com/developer/guide-importing-data-and-etl/ . Este caso se aborda también en el documento de ejercicios titulado “Transformación de una base de datos relacional a un modelo en grafo” pero a un nivel más superficial. Se recomienda leer dicho documento antes de realizar este ejercicio.*

*Se recomienda leer con atención la página web, ya que se introducen conceptos y buenas prácticas a seguir en ‘production’.*

*En la máquina virtual ya hay una base de datos precargada (twitter) que no es necesaria para realizar los ejercicios de esta práctica. Los nodos y relaciones que hay creados no interfieren en el ejercicio, los podéis ignorar tranquilamente.*

*Encontraréis las instrucciones para arrancar el servicio de Neo4j en la página 11 del documento con nombre “Uso de máquina virtual\_ Bases de datos no convencionales.pdf”. Recordad que el usuario / contraseña para acceder a Neo4j son: neo4j / uoc.*

## Ejercicio 3.1 (no puntúa)

Mirar el enunciado de la práctica (carga de los datos).

## Ejercicio 3.2 (30%)

*Se propone responder las siguientes cuestiones:*

*1. Una vez realizada la carga de datos, utilizando cypher y su interfaz web, se pide:*

*a. obtener una visualización del esquema de los nodos/relaciones que se han creado.*

*b. contar cuántos nodos de tipo Supplier se han creado (adjuntar una captura de pantalla de las query y resultado).*

*2. Explicar brevemente (máx 3 líneas) la funcionalidad implementada de la línea:*

*ON CREATE SET c.categoryName = row.CategoryName, c.description = row.Description;*

*3. En la página web, se crea un índice para cada tipo de nodo y una restricción de unicidad para los nodos de tipo Order. Explicar brevemente (máx 3 líneas): ¿Por qué se recomienda crear un índice para cada tipo de nodo?*

### Comprobaciones de la carga de datos

#### Obtener una visualización del esquema

Para obtener una visualización del esquema de los nodos/relaciones que se han creado usamos “*call db.schema,visualization()*”.



Ilustración 10 – función para obtener el esquema.

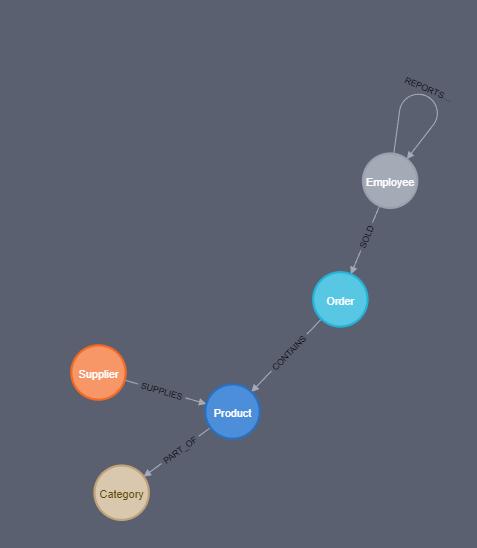


Ilustración 11 - Visualización del esquema.

#### Cuántos nodos tipo *Supplier* se han creado

Para obtener el número de nodos hay que ejecutar la siguiente consulta y obtendremos dicho resultado:



Ilustración 12 - Consulta para obtener el número de nodos de tipo Supplier.

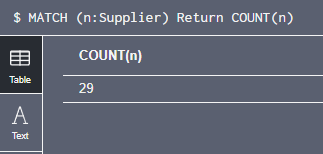
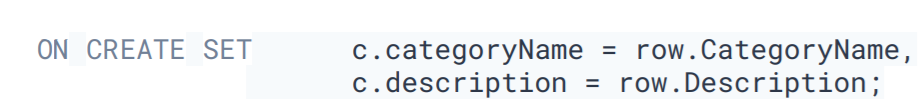


Ilustración 13 - Resultado del número de nodos de tipo Supplier.

### Explicar la funcionalidad implementada



Lo que se está haciendo es definir los valores de las dos propiedades que tiene cada nodo del tipo *Category*, es decir, por cada línea del fichero CSV leído, coge los valores de las columnas (*CategoryName* y *Description*) y los añade a las propiedades correspondientes.

### ¿Por qué se recomienda crear un índice para cada tipo de nodo?

Básicamente se recomienda crear un índice para cada tipo de nodo para garantizar la búsqueda de nodos de forma óptima, dando así un valor único para cada nodo del mismo tipo.

## Ejercicio 3.3 (30%)

*Se pide borrar todos los nodos de tipo Supplier: Adjuntar una captura de pantalla con las queries utilizadas para eliminar todos estos nodos. Ayuda: no es posible borrar nodos que tengan relaciones con otros.*

Para resolver este ejercicio se puede hacer de dos formas, la primera es eliminando primero las relaciones y luego los nodos de tipo *Supplier* (2 consultas), y la segunda es eliminar directamente los nodos pero haciendo uso de la palabra reservada *DETACH* (elimina primero las relaciones y luego los nodos pero todo de una).

En este caso, como no se indica que no se pueda hacer uso de *DETACH*, se ha elegido esta opción ya que es una consulta menos que hay que hacer, dando lugar al siguiente resultado:



Ilustración 14 - Borrado de los nodos tipo Supplier.

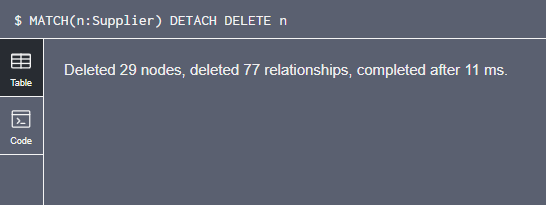


Ilustración 15 - Resultado del borrado de los nodos tipo Supplier.

## Ejercicio 3.4

*Se pide recrear los nodos Supplier y las relaciones SUPPLIES para restablecer el estado de la BBDD y poder responder a las siguientes preguntas:*

*- ¿Cuántos Supplier provienen de Francia? (ayuda: buscar ‘France’ ya que la información es en inglés).*

*- ¿Cuáles son los Supplier cuya web de referencia apunta a una URL absoluta por cada país? Es decir, que contenga una página que empiece por “http”. Se pide identificar cuántos Supplier hay por cada país con una URL absoluta.*

*Se deberá proporcionar las sentencias de creación utilizadas, las dos consultas planteadas y los resultados de las mismas.*

### Recreación de los nodos Supplier y las relaciones SUPPLIES

Lo primero de todo es crear los nodos de tipo *Supplier*, para ello hay que almacenar nuevos campos (*Country, HomePage*) para así satisfacer luego las consultas:

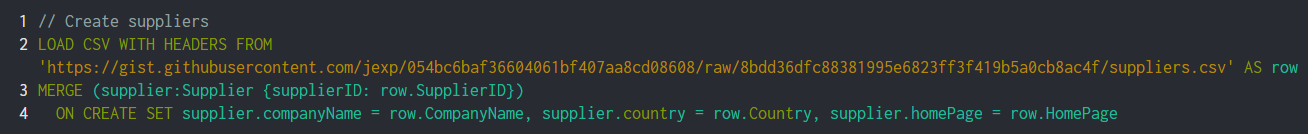


Ilustración 16 - Creación de los nodos de tipo Supplier.

Una vez creado los nodos, creamos las diferentes relaciones, solo se producen con producto:

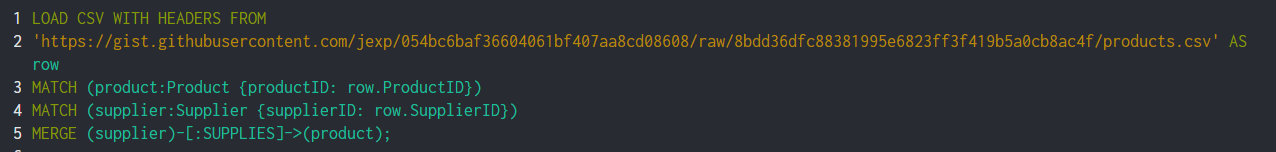


Ilustración 17 - Creación de las relaciones entre Supplier y Product.

### ¿Cuántos Supplier provienen de Francia?

La consulta que tenemos que ejecutar es:



Ilustración 18 - Consulta cuántos Supplier provienen de Francia.

Proporciona el siguiente resultado:

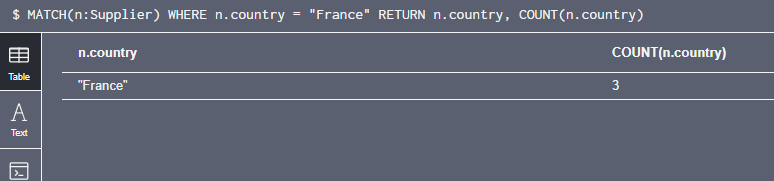


Ilustración 19 - Resultado cuántos Supplier provienen de Francia.

### ¿Cuántos Supplier hay por cada país con una URL absoluta?

La consulta a ejecutar es:



Ilustración 20 - Consulta cuántos países con URL absoluta.

Esta consulta proporciona el siguiente resultado:

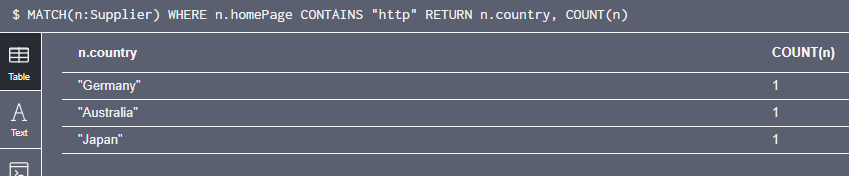


Ilustración 21 - Resultado cuántos países con URL absoluta.

|  |
| --- |
| Ejercicio 4: Neo4j |

*Este ejercicio asume que se ha resuelto el ejercicio anterior. En caso que hayas resuelto el ejercicio 3.B (borrar nodos Supplier) pero no hayas recuperado el estado inicial del DDBB (resuelto el ejercicio 3.C), antes de seguir con este ejercicio crea de nuevo los nodos Supplier y las relaciones SUPPLIES.*

*Se pide proporcionar las siguientes consultas en Cypher y una captura de pantalla con los resultados que se obtienen para las siguientes operaciones:*

## Consulta 1 (20%)

*Encontrar el empleado con título “Sales Representative” que ha vendido (SOLD) más pedidos que contienen la palabra “White”. Listar nombre y apellido del empleado y número de pedidos vendidos.*

La consulta es la siguiente:

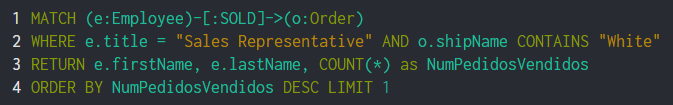


Ilustración 22 – Neo4j consulta 1.

Resultado de la consulta:

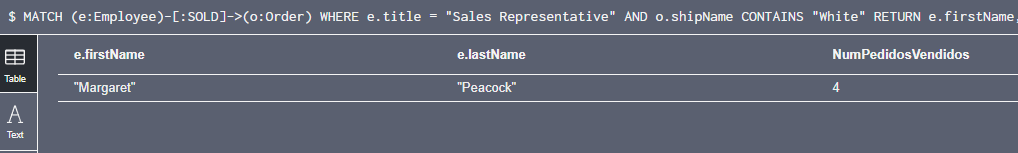


Ilustración 23 - Neo4j resultado consulta 1.

## Consulta 2 (20%)

*Listar el nombre de la categoría asociada (PART\_OF) a 6 productos.*

La consulta realizada es:

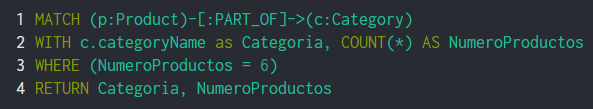


Ilustración 24 - Neo4j consulta 2.

El resultado es:

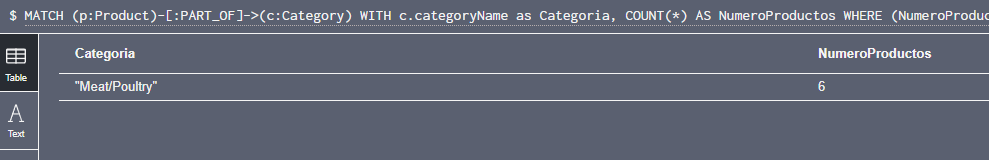


Ilustración 25 - Neo4j resultado consulta 2.

## Consulta 3 (30%)

*Obtener el segundo, tercer y cuarto mejores vendedores de “Tokyo Traders”. Los mejores vendedores son aquellos que han vendido más productos de “Tokyo Traders”. Listar sólo el nombre y apellido de los vendedores y el número de unidades vendidas por vendedor.*

La consulta es:

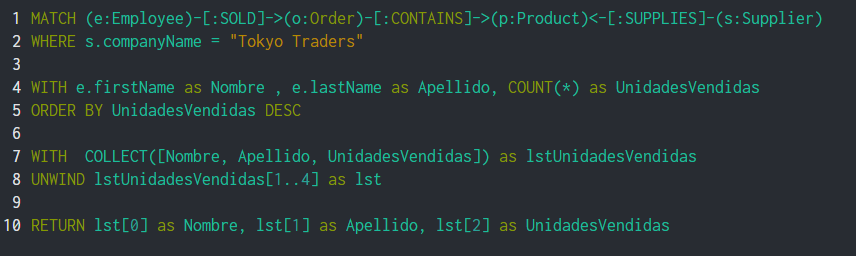


Ilustración 26 - Neo4j consulta 3.

El resultado obtenido:

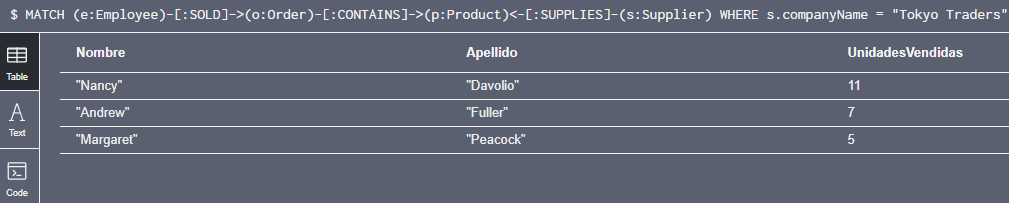


Ilustración 27 - Neo4j resultado consulta 3.

## Consulta 4 (30%)

*El proprietario de las empresas "Leka Trading", "Karkki Oy", "Pavlova, Ltd." quiere centrarse en el mercado del marisco (“Seafood” en la base de datos). Para ello, ha adquirido las empresas "Lyngbysild", "Ma Maison", "Tokyo Traders”. Después de hacerlo, el propietario nos pide que calculemos cuál ha sido el impacto de dichas adquisiciones en el número de productos de marisco vendidos. Para dar respuesta a esta pregunta deberemos calcular el marisco vendido por las empresas originales (antes de la adquisición) y el marisco vendido por todas sus empresas después de su adquisición.*

La consulta es:

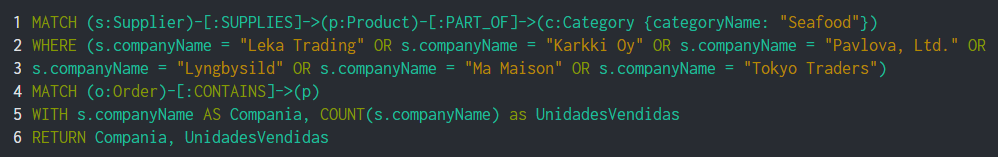


Ilustración 28 - Neo4j consulta 4.

El resultado de la consulta es:

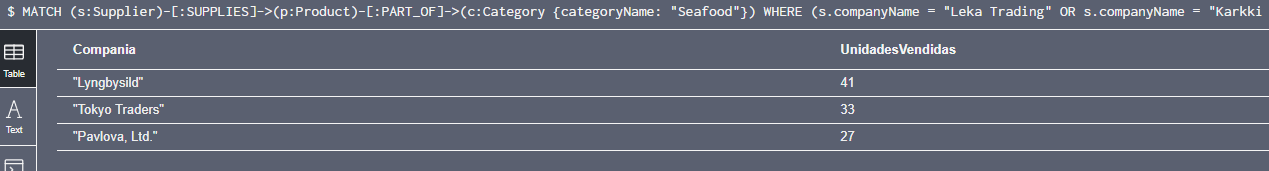


Ilustración 29 - Neo4j resultado consulta 4.

|  |
| --- |
| Bibliografía |

A