



DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico #2

16 de Septiembre de 2016

Base de Datos

Integrante	LU	Correo electrónico
Pedro Rodriguez	197/12	pedro3110.jim@gmail.com
Lucas Tavoraro Ortiz	322/12	tavo92@gmail.com



Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Ciudad Universitaria - Pabellón I

Intendente Güiraldes 2160 - C1428EGA

Ciudad Autónoma de Buenos Aires - Argentina

Tel/Fax: (54 11) 4576-3359

<http://exactas.uba.ar>

Índice

1. Introducción	3
2. Documentos para Mongo	3
2.1. Crímenes cometidos por persona	3
2.1.1. Insert	3
2.1.2. Mantenimiento	3
2.1.3. Map-Reduce	3
2.2. Ciudades con mayor número de crímenes	3
2.2.1. Insert	4
2.2.2. Mantenimiento	4
2.2.3. Map-Reduce	4
2.3. Personas como testigos en casos	4
2.3.1. Insert	4
2.3.2. Mantenimiento	4
2.3.3. Map-Reduce	4
2.4. Casos con mayor número de involucrados o evidencias	5
2.4.1. Insert	5
2.4.2. Mantenimiento	5
2.4.3. Map-Reduce	5
2.5. Crímenes cometidos en los últimos 45 días	5
2.5.1. Insert	6
2.5.2. Mantenimiento	6
2.5.3. Map-Reduce	6
3. Implementación de Map-Reduce	6
4. Conclusiones	6

1. Introducción

En el presente trabajo llevamos nuestro proyecto a MongoDB, analizaremos maneras de representarlo para distintas queries con sus ventajas y desventajas y escribiremos el código de cada map-reduce.

2. Documentos para Mongo

A continuación vamos a definir el documento para Mongo en cada una de las queries dadas y argumentar la elección de dicho documento contando, a alto nivel, como serían los inserts, mantenimiento y búsqueda con map-reduce.

2.1. Crímenes cometidos por persona

1. Para mayor nro. de crímenes cometidos por alguna persona 2. Para nro de crímenes promedio cometidos por criminales

Analizaremos la query que nos pide el mayor número de crímenes cometidos por alguna persona y número de crímenes promedio cometidos. Elegimos diseñar el documento expuesto a continuación:

```
1      [
2          {
3              dniCulpable: 37206752,
4              idCaso: 2
5          }
6      ]
```

Guardaremos en cada documento el **dniCulpable** de la persona que fue encontrada culpable del crime **idCaso**.

2.1.1. Insert

Cada vez que una persona es hallada culpable de un nuevo caso, creamos un nuevo documento con su dni y el caso en cuestión. Esto es una operación muy rápida pues no depende de otros valores.

2.1.2. Mantenimiento

Las ventajas de esta manera de representar la información están en su mantenimiento. Cada documento ocupa poco espacio y, al no depender de nada más, es facil de mantener y escalar.

En un principio analizamos la posibilidad de tener un arreglo de casos por cada dni, pero esto requería mantener una estructura que no sabemos cuanto puede crecer en el tiempo y puede generar problemas al alcanzar el limite de disco. Por eso decidimos utilizar algo más sencillo, pero que podamos garantizar facilmente su mantenimiento y escalabilidad.

2.1.3. Map-Reduce

El map nos debe calcular para cada dni, cuantos casos fueron cometidos por esa persona, o en otras palabras, cuantos documentos hay con ese dni. El reduce depende de la query, en una buscaremos el máximo y en otra el promedio de los valores calculados en el map.

2.2. Ciudades con mayor número de crímenes

Analizaremos la query que nos pide las 10 ciudades con mayor número de crímenes. Elegimos el documento expuesto a continuación:

```
1      [
2          {
3              idCiudad: 4,
4              dniCulpable: 37206752
5          }
6      ]
```

Guardaremos en cada documento, una referencia a la ciudad en **idCiudad** y un culpable de algún crimen como **dniCulpable**.

2.2.1. Insert

Cada vez que se reporte un nuevo crimen, creamos un documento nuevo con la información.

2.2.2. Mantenimiento

Es similar a la anterior querie analizada.

2.2.3. Map-Reduce

El map nos debe calcular para ciudad, cuantos crímenes tiene (idea similar a la querie anterior). El reduce, debe poder ordenarnos la ciudad según sus crímenes y quedarnos con las 10 mas grandes.

2.3. Personas como testigos en casos

La querie consiste en responder cuántas personas se han visto involucradas como testigos en el mayor número de casos. Construimos el siguiente documento:

```
1  [
2      {
3          dniTestigo: 27893435,
4          idCasosInvolucradaComoTestigo:
5              [
6                  {idCaso: 1},
7                  {idCaso: 4},
8                  {idCaso: 6}
9              ]
10     }
11 ]
```

En **dniTestigo** representamos de que testigo se trata y luego tenemos un arreglo denominado **idCasosInvolucradaComoTestigo** en donde cada elemento es un caso en donde esa persona es testigo.

Si bien podíamos optar por un diseño similar a los anteriores, decidimos probar con otro que tiene otras ventajas y desventajas que marcaremos.

2.3.1. Insert

El Insert es mas costoso que en los anteriores pues debemos buscar el documento que contenga el dni del testigo del nuevo caso. Notemos que en caso de no existir un documento debemos crear uno nuevo. Una vez que lo tenemos, debemos agregar un nuevo caso al arreglo.

2.3.2. Mantenimiento

Mencionamos este caso anteriormente. La desventaja que tiene es que va a ir creciendo de tamaño cada documento y eventualmente puede llenar el disco. La solución que se nos ocurre es que, en caso de pasar eso, creamos un nuevo documento para ese testigo en donde a partir de ese momento iremos agregando los casos. En el map-reduce asumiremos que tenemos este diseño ya que es más interesante.

2.3.3. Map-Reduce

El map debe decirnos, para un testigo, en cuantos casos estuvo. Esto es simplemente la suma del largo del arreglo **idCasosInvolucradaComoTestigo** en cada documento en el que aparezca el dni. El reduce debemos primero determinar cuál es que apareció en más casos y luego retornar todas las personas que hayan aparecido en exactamente esa cantidad de casos que calculamos.

Este Map-Reduce es más eficiente que con el diseño usado anteriormente pues recorre muchísimos menos documentos. Recordemos que antes existía un documento por caso, y acá cada documento va a seguir creciendo hasta que no quede más lugar en el disco.

2.4. Casos con mayor número de involucrados o evidencias

Para esas 2 queries utilizaremos un mismo diseño que es el siguiente:

```
1  [
2      {
3          idCaso: 4,
4          {
5              idEvidencias:
6                  [
7                      {idEvidencia: 3},
8                      {idEvidencia: 4}
9                  ],
10             dniInvolucrados:
11                 [
12                     {dniInvolucrado: 37206752}
13                 ]
14         }
15     }
16 ]
```

En donde usaremos **idCaso** como clave, y para cada uno tendremos las distintas evidencias en **idEvidencias** y los involucrados en **dniInvolucrados**.

2.4.1. Insert

A lo largo del tiempo, irán apareciendo evidencias nuevas y gente involucrada. Como en la anterior query, debemos actualizar el documento agregando dicha información según que caso se trate.

2.4.2. Mantenimiento

A diferencia del anterior, en este caso podemos pensar que es realmente muy difícil que llenemos un disco con este documento pues el orden de magnitud de las evidencias y los involucrados no parece ser tanto. En caso de llevar acabo un sistema así, siempre se puede consultar un límite para poder tomar una decisión de diseño de estas características. Además de esta diferencia, en caso de llenarse el disco podemos utilizar una técnica similar a la ya expuesta en la query anterior.

2.4.3. Map-Reduce

Dependiendo la query, dado un caso deberemos contar evidencias o involucrados. Este proceso es análogo al expuesto en el punto anterior, en donde en el map sumaremos el largo de cada arreglo que este relacionado y en el reduce buscaremos el máximo entre todos estos valores calculados.

2.5. Crímenes cometidos en los últimos 45 días

La última query nos pide la cantidad de crímenes de cada tipo cometidos en los últimos 45 días. Elegimos el siguiente diseño para el documento:

```
1  [
2      {
3          idTipoCrimen: 4,
4          cantCrímenesPorFecha:
5              {
6                  {fecha: 2014-10-08,
6                      cantidadDeCrímenes: 4},
```

```

7      {fecha: 2015-03-10,
8        cantidadDeCrímenes: 1},
9      {fecha: 2016-10-10,
10       cantidadDeCrímenes: 2}
11    ]

```

En donde **idTipoCrimen** es la clave que representa un tipo de crimen, y el mapa **cantCrímenesPorFecha**, cuyas claves son fechas y los valores cantidad de crímenes en cada una. Este documento tendrá la propiedad de que solo mantendremos fechas en los últimos 45 días guardadas (En Mantenimiento desarrollaremos mejor este punto).

2.5.1. Insert

Buscamos el tipo de crimen correspondiente para tener el documento (o crear un nuevo si no existe). Luego, buscamos en el diccionario la clave correspondiente a la fecha, e incrementamos su valor. En caso de no existir dicha clave, la creamos con el valor 1.

2.5.2. Mantenimiento

Para optimizar el uso del disco, creamos un documento para cada tipo de crimen, y dicho documento tiene una cantidad de espacio acotado pues puede contener como mucho 45 fechas distintas. Esto es una ventaja a la hora de escalar esta estructura.

Para poder garantizar la propiedad de que solo contenga crímenes de los últimos 45 días, cuando cambia de día debemos realizar una tarea de mantenimiento. Hay que recorrer el diccionario correspondiente a cada tipo de crimen, y borrar las entradas correspondientes a fechas de hace más de 45 días. Este proceso puede ser rápido, pues no creemos que existan muchísimos tipos de crímenes y como mencionamos, hay como mucho 45 fechas. Esto quiere decir que en este proceso eliminaremos como mucho 1 de ellas, ya que al cambia de día alguna puede haber pasado a ser 46 días antes.

2.5.3. Map-Reduce

En el map tendremos el tipo de crimen como la clave, y calcularemos la la suma de todos los crímenes en las fechas que hay guardadas (Recordemos que todas están incluidas en los últimos 45 días). El reduce es simplemente tomar estos valores y mostrarlos.

3. Implementación de Map-Reduce

4. Conclusiones