UNIVERSIDAD DE GRANADA

Cloud Computing

Servicios de bases de datos en *cloud*: MongoDB

Práctica 3

Víctor Vázquez Rodríguez
victorvazrod@correo.ugr.es
76664636R

Máster universitario en Ingeniería Informática

Curso 2019/20

1. Introducción

En esta práctica, el objetivo es realizar consultas sobre una base de datos MongoDB para obtener la información especificada en el guión. La base de datos con la que se va a trabajar se compone de crímenes que se registraron en la ciudad estadounidense de Sacramento en enero de 2006.

Para la realización de esta práctica he hecho uso de MongoDB en mi ordenador local mediante Docker, por lo que he dedicado la siguiente sección a explicar brevemente el proceso de configuración de este entorno.

2. Configuración

Para lanzar el contenedor Docker de MongoDB, lo único que tenemos que ejecutar es `docker run -d --name <contenedor>mongo`. Aunque se puede especificar un usuario y contraseña para la base de datos mediante el paso de variables de entorno al contenedor, para el desarrollo de esta práctica no son necesarias estas medidas de seguridad.

Ahora, copiamos el fichero CSV que contiene los datos dentro del contenedor con `docker cp <fichero>.csv <contenedor>:/tmp/data.csv`, lo cuál creará un nuevo archivo dentro del contenedor en `/tmp/data.csv` con dichos datos.

Por último, cargamos los datos del CSV a la base de datos Mongo con la herramienta `mongoimport` con el comando que se muestra a continuación, donde *bdd* y *coleccion* son los nombres de la base de datos y la colección donde insertar los datos, respectivamente.[1]

```
docker exec -it <contenedor> mongoimport -d <bdd>
    -c <coleccion> --type csv --file /tmp/data.csv
    --headerline
```

3. Consultas

3.1. Contar el número de delitos/robos

En nuestra colección, todos los documentos representan un delito, por lo que para obtener el número de delitos solo tenemos que obtener el número de documentos, lo cuál podemos hacer con la siguiente consulta:

```
db.crimes.find().count()
```

Hay 7584 delitos registrados en la colección.

[1] En mi caso, el nombre de esta base de datos es *ccsa* y, el de la colección, *crimes*.

3.2. Contar el número de delitos por hora

Los documentos de la colección poseen un campo *cdatetime* con la fecha y la hora del crimen. Como este campo no está en un formato estándar de fecha y hora, debemos usar una expresión regular para extraer la hora. Por ello, en nuestra agregación incorporamos un primer paso o *step* en el que se aplica una expresión regular a la fecha con `$regexFind` y se guarda el valor en un nuevo campo *hour* con `$set`.

Una vez hecho esto, se agrupan los documentos por la hora con `$group` y se calcula el *count* con `$sum`. La consulta completa es la siguiente:

```
db.crimes.aggregate([
    {
        $set: {
            hour: {$regexFind: {
                input: "$cdatetime",
                regex: /\d{1,2}(?=:)/
            }}
        }
    },
    {
        $group: {
            _id: "$hour.match",
            count: {$sum: 1}
        }
    }
])
```

El resultado de realizar la consulta también se puede ver a continuación, donde *_id* contiene el valor de la hora en cuestión.

```
{ "_id" : "10", "count" : 306 }
{ "_id" : "4", "count" : 91 }
{ "_id" : "6", "count" : 114 }
{ "_id" : "2", "count" : 171 }
{ "_id" : "16", "count" : 471 }
{ "_id" : "17", "count" : 458 }
{ "_id" : "11", "count" : 338 }
{ "_id" : "3", "count" : 122 }
{ "_id" : "21", "count" : 358 }
{ "_id" : "18", "count" : 430 }
{ "_id" : "9", "count" : 326 }
{ "_id" : "22", "count" : 384 }
{ "_id" : "23", "count" : 304 }
{ "_id" : "19", "count" : 358 }
{ "_id" : "8", "count" : 380 }
{ "_id" : "20", "count" : 362 }
```

```
{ "_id" : "13", "count" : 353 }
{ "_id" : "15", "count" : 417 }
{ "_id" : "0", "count" : 578 }
{ "_id" : "7", "count" : 219 }
{ "_id" : "14", "count" : 385 }
{ "_id" : "12", "count" : 399 }
{ "_id" : "1", "count" : 187 }
{ "_id" : "5", "count" : 73 }
```

3.3. Mostrar los 5 delitos más habituales

Los documentos tienen un campo *ucr_ncic_code* que contiene un código identificativo del crimen cometido. Por lo tanto, para obtener los 5 delitos más habituales solo tendríamos que agrupar los documentos por este código, realizar un *count* como en la consulta anterior y, además, ordenar los resultados y quedarnos con los 5 primeros. Estos dos últimos pasos se realizan con `$sort` y `$limit`, respectivamente. Para poder ver también de qué delito se trata, se ha añadido la descripción del mismo (*crimedescr*) al resultado con `$addToSet`.

```
db.crimes.aggregate([
    {
        $group: {
            _id: "$ucr_ncic_code",
            crimedescr: {
                $addToSet: "$crimedescr"
            },
            count: {$sum: 1}
        }
    },
    {
        $sort: {count: -1}
    },
    {
        $limit: 5
    }
])
```

La respuesta de esta consulta la encontramos a continuación:

```
{
    "_id" : 7000,
    "crimedescr" : [
        "ACCIDENTAL FIRES/ARSON -I RPT",
        "FOUND PROPERTY - I RPT",
        "RPT # CANCELLED- I RPT",
        "VICE/GAMBLING ACT - I RPT",
```

```
"JUVENILE DISURBANCE - I RPT",
"VANDALISM - I RPT",
"POSSIBLE FINANCIAL CRIME-I RPT",
"FALSE PERSONATION - I RPT",
"THREATS - I RPT",
"TOWED/STORED VEH-14602.6",
"NEIGHBORHOOD DISTURBANCE-I RPT",
"RECOVERED PROPERTY - I RPT",
"MISSING PERSON LOCATE O/S ASSI",
"POSS STOLEN VEHICLE- I RPT",
"ABC LICENSE VIO/INFO - I RPT",
"TOWED/STORED VEHICLE",
"LOST PROPERTY - I RPT",
"CASUALTY REPORT",
"MISCELLANEOUS I RPT (ZMISC)",
"TERRORIST THREATS - I RPT",
"MISSING PERSON I RPT",
"PROTECTIVE CUSTODY-I RPT",
"GANG ACTIVITY - I RPT",
"NARCOTICS SUSP/EVID/ACT- I RPT",
"WANTED SUBJ-O/S WANT/ I RPT",
"ABANDONED VEHICLE - I RPT",
"601(A) WI JUVENILE INCORRIGIBL",
"TRESPASS OR PROWLER- I RPT",
"BURGLARY - I RPT",
"LOST PROPERTY-MATRICULA I RPT",
"FAMILY DISTURBANCE - I RPT",
"5150 WI DANGER SELF/OTHERS",
"BOMBS/THREATS/EXPLOSIV- I RPT",
"JUVENILE PROBLEMS - I RPT",
"PETTY THEFT - I RPT",
"HAZARDOUS SITUATION - I RPT",
"IMPOUNDED VEHICLE",
"O/S AGENCY -ASSISTANCE- I RPT",
"NON INJ HR/MAIL OUT REPORT",
"HOMICIDE ASSAULT - I RPT",
"PERSON INFORMATION - I RPT",
"CHILD WELFARE - I RPT",
"DUI I RPT",
"DAMAGE - I RPT",
"POSSIBLE MENTAL - I RPT",
"3056 PAROLE VIO - I RPT",
"BATTERY - I RPT",
"FILE-UNABLE TO DEFINE - I RPT",
"BUSINESS PERMITS - I RPT",
"POSSIBLE STOLEN PROPERTY-I RPT",
```

```
            "CHILD CUSTODY - I RPT",
            "849(B)(1) CERTIFICATE OF RELEA",
            "INTOX REPORT/ADMIN PER - I RPT",
            "WARRANT SERVED - I RPT",
            "SUSP PERS-NO CRIME - I RPT",
            "HARASSMENT - I RPT",
            "MISSING PERSON",
            "FRAUDULENT DOCUMENTS- I RPT",
            "CAR CLOUT -  I RPT",
            "ELDER ABUSE/PHYS/MENTAL-I RPT",
            "TRAFFIC - I RPT",
            "GRAFFITI PROBLEMS- I RPT",
            "SHOOT INTO OCCUP DWELL - I RPT",
            "SUSPICIOUS VEHICLE - I RPT",
            "GRAND THEFT - I RPT",
            "PROB VIOLA/FEL-MISD - I RPT",
            "ROBBERY - I RPT",
            "FOUND EVIDENCE-NON NARC -I RPT",
            "FRAUD OR BUNCO - I RPT",
            "SAFEKEEPING - I RPT",
            "TELEPEST -I RPT",
            "HIT AND RUN /SUSPECTS- I RPT",
            "TRAFFIC ARREST FOR DA-I RPT",
            "ASSAULT WITH WEAPON - I RPT"
        ],
        "count" : 2470
    }
    {
        "_id" : 2404,
        "crimedescr" : [
            "10851(A)VC TAKE VEH W/O OWNER",
            "10851 VC AUTO THEFT LOCATE",
            "10855 VC LSE/RENT NOT RETURNED",
            "487(D) PC GRAND THEFT AUTO",
            "BAIT CAR 10851 VC TAKE VEHICLE"
        ],
        "count" : 881
    }
    {
        "_id" : 2299,
        "crimedescr" : [
            "459 PC  BURGLARY VEHICLE",
            "459 PC  BURGLARY-UNSPECIFIED"
        ],
        "count" : 474
    }
```

```
{
    "_id" : 5400,
    "crimedescr" : [
        "TRAFFIC-ACCIDENT-NON INJURY",
        "TRAFFIC-ACCIDENT INJURY"
    ],
    "count" : 357
}
{
    "_id" : 2999,
    "crimedescr" : [
        "594.3(A)VANDAL/PLACE OF WORSHI",
        "10852 VC VEHICLE TAMPERING",
        "1705 US DESTRUCT OF MAIL RECEP",
        "594(B)(2)(A) VANDALISM/ -$400",
        "594(B)(1)VANDALISM GRAF +$400",
        "594(B)(2)(A)VANDALISM GRAF-400",
        "594(B)(1)PC  VANDALISM +$400",
        "603  FORCED ENTRY/PROP DAMAGE",
        "10853 VC MALIC MISCHIEF TO VEH"
    ],
    "count" : 356
}
```

Como se puede ver, todos los códigos agrupan a más de una descripción de crimen, en algunos casos, siendo crímenes diferentes. Podríamos entonces modificar la consulta, agregando por descripción del crimen, para obtener un resultado mas real de cuáles son los crímenes más habituales.

```
db.crimes.aggregate([
    {
        $group: {
            _id: "$crimedescr",
            count: {$sum: 1}
        }
    },
    {
        $sort: {count: -1}
    },
    {
        $limit: 5
    }
])
```

En el resultado, vemos como se trata sobre todo de delitos relacionados con vehículos excepto el último, que es de robos en viviendas.

```
{
    "_id" : "10851(A)VC TAKE VEH W/O OWNER",
    "count" : 653
}
{
    "_id" : "TOWED/STORED VEH-14602.6",
    "count" : 463
}
{
    "_id" : "459 PC  BURGLARY VEHICLE",
    "count" : 462
}
{
    "_id" : "TOWED/STORED VEHICLE",
    "count" : 434
}
{
    "_id" : "459 PC  BURGLARY RESIDENCE",
    "count" : 356
}
```