BIG DATA I
MÁSTER EN CIENCIA DE DATOS E INGENIERIA DE COMPUTADORES

JAUME CLOQUELL CAPO
jaumecloquell@correo.ugr.com

Database

La base de datos usada puede ser consultada y descargada en catalog. Este conjunto de datos refleja los incidentes de crimen reportados (con la excepción de los asesinatos en los que existen datos para cada víctima) que ocurrieron en la Ciudad de Chicago desde el 2001 hasta el presente, menos los siete días más recientes.

Está formada por 6822518 filas y 22 columnas (ID,Case Number,Date,Block,IUCR,Primary Type,Description,Location Description,Arrest,Domestic,Beat,District,Ward,Community Area,FBI Code,X Coordinate,Y Coordinate,Year,Updated On,Latitude,Longitude,Location) de distintos formatos y con valores NULL en algunas de ellas.

Creación de la base de datos

Para poder generar las consultas sin interferir con las demás que se realicen en el cluster crearemos una base de datos propia.

Para ello,una vez dentro del cluster de Hadoop, primero conectamos a impala-shell:

```
impala-shell
```

Tras esto, creamos la tabla mediante el siguiente script sql. Crimes_-_2001_to_present.csv contiene en la primera row, el nombre de las columnas. Por tanto, se deberá evitar su lectura al importarlo en la tabla crimes mediante el siguiente script.

```
CREATE DATABASE CRIMES IF NOT EXIST
USE DATABASE CRIMES
CREATE TABLE IF NOT EXISTS crimes (
   id INT,
   case_number STRING,
   fecha STRING,
   block STRING,
   iucr STRING,
   primary_type STRING,
   description STRING,
   location_description STRING,
   arrest BOOLEAN,
   domestic BOOLEAN,
   beat INT,
   district INT,
   ward INT,
   community_area STRING,
   fbi_code INT,
   x_coordinate INT,
   y_coordinate INT,
   year INT,
   updated_on STRING,
   latitude STRING,
   longitude STRING,
   localizacion STRING
) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED
AS TEXTFILE;
alter table crimes set tblproperties('skip.header.line.count'='1');
```

Carga de Datos

Una vez creada la base de datos y la tabla debemos cargar los datos en la misma, para ello, primero deberemos mandar los datos al HDFS de hadoop.

```
hdfs dfs -mkdir /user/impala/prueba
hdfs dfs -put /home/cloudera/Downloads/Crimes_-_2001_to_present.csv
/user/impala/prueba.
```

Si todo a ido bien, observaremos que la base de datos ahora está en el hdfs de hadoop, por lo que podremos cargar los datos en una tabla, para ello cargamos el impala-shell y usamos (comando USE) nuestra base de datos creada anteriormente.Tras esto cargamos los datos:

```
LOAD DATA INPATH '/user/impala/prueba/Crimes_-_2001_to_present.csv'
OVERWRITE INTO TABLE crimes;
```

Experimento de datos

Para la resolución del trabajo, se ha relacida una query que contiene todos los aspectos definidos en el enunciado del presente trabajo:

- Debe incluir una operación de proyección.
- Debe incluir una operación de selección.
- Debe incluir agrupamientos (group) y resúmenes de información.

```
SELECT primary_type, Count(*) as TheCount
FROM
(
  SELECT *
  FROM crimes WHERE year
  BETWEEN 2010 AND 2019
) sub
GROUP BY primary_type;
```

La query esta estructura en una query padre que trabaja a partir de una subquery. Para la descripción del script iremos de dentro (subquery) hacia fuera. En primer lugar, la subquery realiza un operación de selección donde filtra sólo los crímenes entre el 2010 y el 2019. Esta subquery devuelve todos los campos ya que la query padre se encargará de proyectar los valores que más le interesen.

La Query padre, a partir de los datos seleccionados en la anterior consultas, agrupa los resultados por el campo “primary_type” y hace una cuenta del número de crímenes por cada grupo.

Por tanto la anterior query, devuelve el número de crímenes, que se han realizado entre el

2010 hasta el día de hoy, agrupados por tipo. La imagen 2, representa el resultado del query anteriormente descrita.

primary_type	thecount
CRIM SEXUAL ASSAULT	13108
OBSCENITY	394
NON - CRIMINAL	38
BURGLARY	160981
ARSON	4035
CRIMINAL DAMAGE	288927
INTERFERENCE WITH PUBLIC OFFICER	10555
KIDNAPPING	1976
DECEPTIVE PRACTICE	140244
PUBLIC INDECENCY	98
MOTOR VEHICLE THEFT	112801
NARCOTICS	240186
HOMICIDE	4827
WEAPONS VIOLATION	34476
OTHER NARCOTIC VIOLATION	44
THEFT	603397
PUBLIC PEACE VIOLATION	20449
INTIMIDATION	1324
OFFENSE INVOLVING CHILDREN	20694
BATTERY	471901
SEX OFFENSE	8934
ROBBERY	107218
CRIMINAL TRESPASS	67054
STALKING	1579
LIQUOR LAW VIOLATION	3701
HUMAN TRAFFICKING	54
CONCEALED CARRY LICENSE VIOLATION	326
PROSTITUTION	14069
OTHER OFFENSE	164637
NON-CRIMINAL (SUBJECT SPECIFIED)	9
GAMBLING	4227
NON-CRIMINAL	169
ASSAULT	164270

Imagen 2. Resultado de la query