**Neo4j**

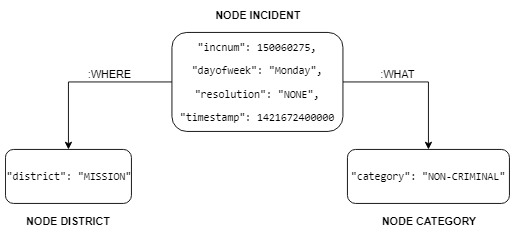
**Documentación Neo4j**

**Grafo creado**

Tras estudiar el dataset proporcionado se optó por implementar un grafo con tres tipos de nodos (***Incident***, ***Category*** y ***District***) y dos tipos de relaciones entre estos nodos: relación ***WHERE*** de ***Incident*** a ***District*** y relación ***WHAT*** de ***Incident*** a ***Category***.

Entrando en detalle en cada tipo de nodo, ***Incident*** incluye los campos *incnum, dayofweek, resoultion* y *timestamp*, que equivalen al número de incidente (id único), el día de la semana en el que tuvo lugar el incidente, la resolución de este y el instante de tiempo en el que ocurrió el crimen, respectivamente. Por su parte, el nodo ***District*** solo incluye el campo *district*, que es el nombre del distrito en cuestión. Por último, ***Category*** cuenta solo con el campo *category*, el tipo de categoría en la que se incluyen los incidentes.

En la Figura 1, podemos observar un ejemplo de un incidente (nodo ***Incident***) relacionándose con el distrito (relación ***WHERE*** hacia un nodo ***District***) y con la categoría (relación ***WHAT*** hacia un nodo ***Category***).



*Figura 1. Ejemplo simple del grafo implementado.*

**Procesamiento del fichero de datos proporcionado**

El problema que nos encontramos de cara a volcar el contenido del dataset en la base de datos de Neo4j es que algunos números de incidente se repiten, por lo que no se podía utilizar este número como identificador único de los nodos ***Incident***.

Por ello, se llevó a cabo un pre-procesado del fichero de datos proporcionado. Este pre-procesado consistió en obtener otros cinco datasets para crear nuestro grafo. Los ficheros de datos generados fueron los siguientes:

* **Districts.csv**: los distritos de San Francisco que aparecen en el dataset original. Son un total de 10.
* **Categories.csv**: las diferentes categorías de crímenes. Un total de 39.
* **Incidents.csv**: todos los incidentes **ÚNICOS** que han tenido lugar en San Francisco. Más de 1.700.000 incidentes.
* **IncToDistrict.csv**: relaciona los incidentes **ÚNICOS** con el distrito en el que tuvieron lugar.
* **IncToCategory.csv**: relaciona los incidentes con la categoría en la que se clasifican. En este caso sí que se pueden repetir los números de incidente, ya que un mismo incidente se puede clasificar en varias categorías.

La generación de estos nuevos dataset se realizó en el proyecto Maven ProcessForNeo, en el que se usa la dependencia OpenCSV, que permite leer y parsear un CSV. Gracias a la mencionada dependencia se obtuvieron los campos necesarios para la construcción de cada uno de los nuevos ficheros de datos ya explicados. El proyecto ProcessForNeo también se encuentra en nuestro repositorio, para su posible revisión y/o estudio.

**Modificación del fichero de configuración de Neo4j y volcado de los datos**

Antes de volcar el contenido de estos ficheros en nuestra base de datos orientada a grafos debemos cambiar una serie de parámetros del fichero de configuración de la BD. Dicho archivo se denomina **neo4j.conf** y se encuentra en la siguiente ruta:

C:\Users\**USUARIO**\.Neo4jDesktop\neo4jDatabases\**DB**\installation-**X.X.X**\conf

Los parámetros en negrita son:

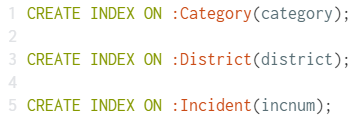
* **USUARIO**: usuario en el que estemos trabajando en nuestra máquina local.
* **DB**: base de datos Neo4j a la cual queremos acceder a su fichero de configuración.
* **X.X.X.**: versión elegida a la hora de crear la base de datos.

Una vez abramos el fichero de configuración debemos comentar cuatro líneas:

* **dbms.directories.import=import**: para poder importar CSVs desde cualquier directorio de nuestro equipo
* **dbms.memory.heap.initial\_size, dbms.memory.heap.max\_size** y **dbms.memory.pagecache.size**: para que la importación de datos, consultas y borrados vayan más rápidos.

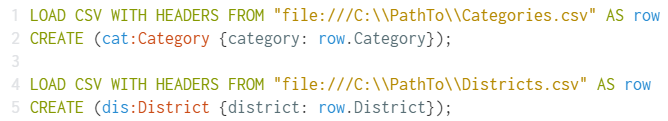
Ahora ya podemos crear los nodos y las relaciones de nuestro grafo. Para ello utilizaremos el browser de la aplicación de escritorio de Neo4j y ejecutamos el script **importNeo.txt** (también se encuentra en nuestro repositorio).

En primer lugar, creamos índices para los campos category (del nodo Category), district (del nodo District) y incnum (del nodo Incident), lo cual nos permitirá realizar búsquedas con mayor facilidad por el grafo.



*Figura 2. Creación de los índices.*

A continuación, procedemos a crear los nodos Category, District e Incident. En la creación de los nodos Incident además usaremos la cláusula **USING PERIODIC COMMIT 5000**, para que se haga un commit cada 5000 nodos creados.



*Figura 3. Creación de los nodos Category y District.*



*Figura 4. Creación de los nodos Incident.*

Por último, crearemos las relaciones de nuestro grafo. Al igual que en la creación de los nodos de incidentes, haremos commit cada 5000 relaciones creadas.



*Figura 5. Creación de las relaciones WHERE de Incident a District.*



*Figura 6. Creación de las relaciones WHAT de Incident a Category.*