



PD03: NoSQL

Integrantes

Basto Martin Manuel Yahir
Buenfil Perez Maricarmen
Rodríguez Alfaro Sergio Daniel

Profesor. Luis Ramiro Basto Diaz

FACULTAD DE MATEMÁTICAS
LICENCIATURA EN CIENCIAS DE LA COMPUTACIÓN
MODELADO DE DATOS
FECHA DE ENTREGA: 3 DE DICIEMBRE DE 2024

INDICE

Obtención del Dataset y Modelado de la Base de Datos	3
Seleccionar y describir el dataset	3
Modelar la base de datos	3
Diccionario de datos	4
Nodos	4
Relaciones	6
Importación de Datos	7
Instalación y Configuración	7
Carga de Datos en Neo4j	7
Referencias	9

Obtención del Dataset y Modelado de la Base de Datos

Seleccionar y describir el dataset

Paradise Papers by ICIJ Graph Example

<https://github.com/neo4j-graph-examples/icij-paradise-papers>

Los datos que une a políticos, delincuentes y una industria corrupta que oculta su dinero

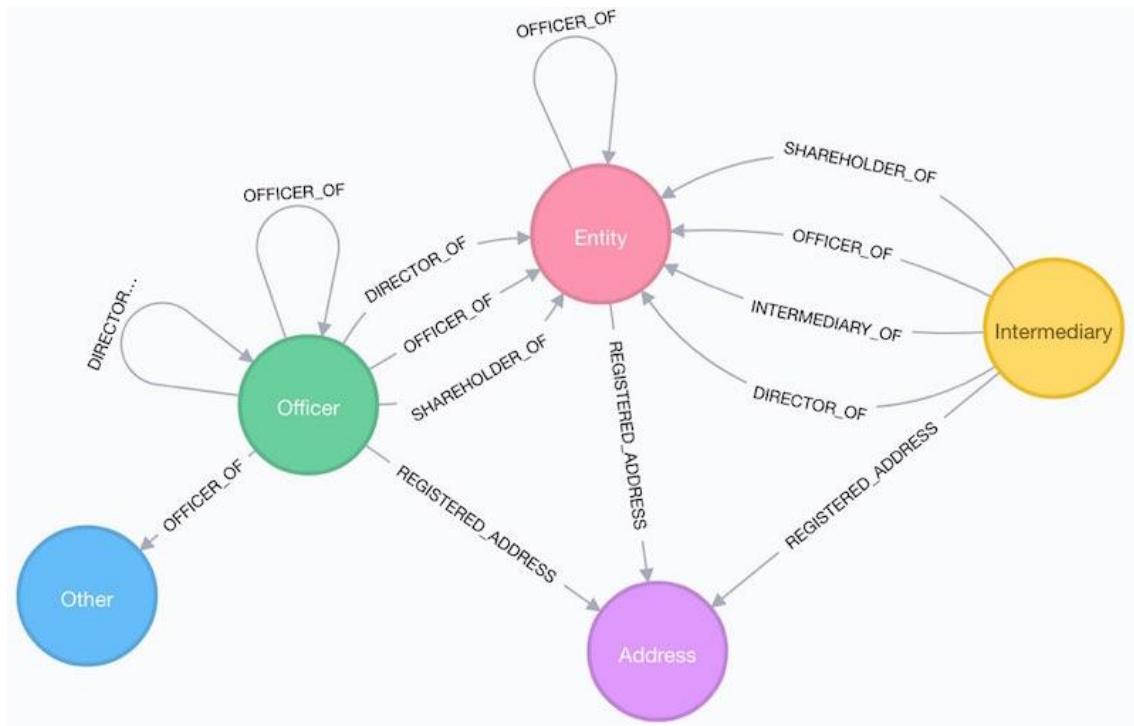
La base de datos *Offshore Leaks* del ICIJ, con la que estás trabajando en Neo4j, contiene información sobre casi 350,000 entidades offshore que forman parte de las investigaciones de los *Papeles de Panamá*, los *Papeles del Paraíso* y las *Offshore Leaks*. Los datos abarcan un largo período de actividades y están vinculados a personas y empresas en más de 200 países y territorios.

El verdadero valor de esta base de datos radica en que elimina el secreto que rodea a las empresas y fideicomisos constituidos en paraísos fiscales, exponiendo a las personas detrás de ellos. Esto incluye, cuando está disponible, los nombres de los verdaderos propietarios de esas estructuras opacas. En total, revela más de 430,000 nombres de personas y empresas detrás de estas estructuras offshore secretas. Los datos provienen de registros filtrados y no de un registro corporativo estandarizado, por lo que pueden existir duplicados. Se sugiere confirmar las identidades de las personas o entidades encontradas en la base de datos utilizando direcciones u otra información identificable.

Existen usos legítimos para las empresas y fideicomisos offshore. No es nuestra intención sugerir ni implicar que las personas, empresas u otras entidades incluidas en la base de datos hayan infringido la ley o actuado de manera inapropiada. Si encuentras algún error en la base de datos, por favor contacta con el ICIJ.

Modelar la base de datos

La base de datos *Offshore Leaks* fue importada a Neo4j para que periodistas e investigadores puedan aprovechar las conexiones presentes en los datos. A la izquierda se muestra el modelo de datos básico de un "grafo de propiedades". Cada registro de datos se representa como un "nodo", que puede corresponder a una entidad, intermediario, oficial o dirección. Estos nodos están interconectados para formar un "grafo" que revela una red compleja de relaciones. A la izquierda puedes ver un diagrama simplificado que muestra cómo los nodos se conectan entre sí.



Los datos de *Offshore Leaks* revelan un conjunto de conexiones entre personas y entidades offshore. Las bases de datos en grafo son la mejor manera de explorar las relaciones entre estas personas y entidades, ya que resultan mucho más intuitivas para este propósito que una base de datos SQL u otros tipos de bases de datos NoSQL.

Diccionario de datos

Nodos

Estos son los tipos de nodos que encontrarás en los datos:

1. **Entidad** (*Entity*) - Una entidad legal offshore. Esto podría ser una empresa, fideicomiso, fundación u otra entidad legal creada en una jurisdicción de baja tributación.
2. **Oficial** (*Officer*) - Una persona o empresa que desempeña un rol en una entidad offshore, como beneficiario, director o accionista. Las relaciones mostradas en el diagrama son solo una muestra de todas las existentes.
3. **Intermediario** (*Intermediary*) - Un intermediario entre alguien que busca crear una corporación offshore y un proveedor de servicios offshore, como una firma de abogados o un agente que solicita a un proveedor que cree una firma offshore.
4. **Dirección** (*Address*) - La dirección registrada tal como aparece en las bases de datos originales obtenidas por el ICIJ.
5. **Otros** (*Other*) - Otras entidades encontradas en los datos.

Cada nodo de intermediario tiene las siguientes propiedades:

Cada nodo de *Intermediary* representa un intermediario entre alguien que busca una corporación offshore y un proveedor de servicios offshore, que generalmente es un bufete de abogados o un mediador que solicita a un proveedor de servicios offshore la creación de una empresa offshore para un intermediario.

- **name** (*nombre*): El nombre del intermediario.
- **address** (*dirección*): La dirección del intermediario.
- **sourceID** (*ID de fuente*): *Offshore Leaks*, *Panama Papers* o *Paradise Papers*, dependiendo de la fuente de los datos.
- **status** (*estado*): [No se especifica en el texto, pero podría ser el estado operativo del intermediario].
- **valid_until** (*válido hasta*): [La fecha hasta la cual es válido].
- **country_codes** (*códigos de países*): Los códigos de países asociados.
- **countries** (*países*): Los nombres de los países asociados.

Cada nodo de entidad tiene las siguientes propiedades:

Cada *Entity* es una empresa, fideicomiso o fondo creado en una jurisdicción offshore de bajos impuestos por un agente.

- **name** (*nombre*): El nombre de la entidad legal.
- **sourceID** (*ID de fuente*): *Offshore Leaks*, *Panama Papers* o *Paradise Papers*, dependiendo de la fuente de los datos.
- **address** (*dirección*): Este campo incluye la dirección registrada asociada con la entidad, pero solo cuando coincide con la del intermediario. De lo contrario, la información de la dirección registrada se almacena en un nodo de dirección conectado a este nodo de entidad mediante una relación de tipo *REGISTERED_ADDRESS*.
- **former_name** (*nombre anterior*), **original_name** (*nombre original*): Nombres previos o originales de la entidad.
- **company_type** (*tipo de compañía*): El tipo de entidad legal.
- **status** (*estado*): El estado actual de la entidad.
- **incorporation_date** (*fecha de incorporación*), **inactivation_date** (*fecha de inactivación*), **struck_off_date** (*fecha de eliminación*), **dorm_date** (*fecha de inactividad*): Fechas relacionadas con eventos en la trayectoria de la compañía.
- **service_provider** (*proveedor de servicios*): El proveedor de servicios offshore asociado con la entidad.
- **ibcRUC**: [Información no especificada en el texto, probablemente un identificador fiscal o de registro].
- **valid_until** (*válido hasta*): Fecha hasta la cual es válida la entidad.
- **jurisdiction** (*jurisdicción*), **jurisdiction_description** (*descripción de la jurisdicción*): Información sobre la jurisdicción en la que se encuentra registrada la entidad.
- **country_codes** (*códigos de países*), **countries** (*países*): Códigos y nombres de los países asociados con la entidad.

Los nodos de oficial tienen las siguientes propiedades:

Cada nodo de *Officer* representa al beneficiario, director o accionista de la entidad legal offshore, como un beneficiario, accionista, director, etc.

- **name** (*nombre*): El nombre del oficial.
- **valid_until** (*válido hasta*): Fecha hasta la cual es válida la información.
- **sourceID** (*ID de fuente*): *Offshore Leaks*, *Panama Papers* o *Paradise Papers*, dependiendo de la fuente de los datos.

- **country_codes** (códigos de países), **countries** (países): Códigos y nombres de los países asociados.

Los nodos de dirección tienen las siguientes propiedades:

El nodo de *Address* representa la dirección registrada en los archivos del Intermediario, Oficial o Entidad.

- **address** (dirección): La dirección tal como aparece en los registros.
- **sourceID** (ID de fuente): *Offshore Leaks*, *Panama Papers* o *Paradise Papers*, dependiendo de la fuente de los datos.
- **valid_until** (válido hasta): Fecha hasta la cual es válida la información.
- **country_codes** (códigos de países), **countries** (países): Códigos y nombres de los países asociados.

Relaciones

En el modelo de datos aparecen las siguientes relaciones:

(:Intermediary)-[:INTERMEDIARY_OF]->(:Entity)

(:Intermediary|Officer|Entity)-[:REGISTERED_ADDRESS]->(:Address)

(:Officer|Intermediary)-[:OFFICER_OF]->(:Entity)

fromLabel	relType	toLabels	frequency
["Officer"]	"OFFICER_OF"	["Entity", "Officer", "Intermediary"]	217385
["Officer"]	"REGISTERED_ADDRESS"	["Address"]	89195
["Entity"]	"REGISTERED_ADDRESS"	["Address"]	38890
["Intermediary"]	"CONNECTED_TO"	["Entity"]	9229
["Other"]	"INTERMEDIARY_OF"	["Entity"]	4063
["Other"]	"OFFICER_OF"	["Entity"]	3724
["Officer"]	"CONNECTED_TO"	["Entity"]	1099

Intermediary

(:Intermediary)-[:INTERMEDIARY_OF]->(:Entity)

(:Intermediary|Officer|Entity)-[:REGISTERED_ADDRESS]->(:Address)

(:Officer|Intermediary)-[:OFFICER_OF]->(:Entity)

Entity

(:Intermediary)-[:INTERMEDIARY_OF]->(:Entity) - La relación que muestra al intermediario que supervisó la creación de la entidad.

(:Intermediary|Officer|Entity)-[:REGISTERED_ADDRESS]->(:Address) - La dirección registrada de la entidad.

(:Entity)-[:RELATED_ENTITY]->(:Entity) - Entidades que, según los documentos filtrados, estaban conectadas entre sí.

Officer

(:Intermediary|Officer|Entity)-[:REGISTERED_ADDRESS]->(:Address)

(:Officer)-[:OFFICER_OF]->(:Entity)

Address

(:Intermediary|Officer|Entity)-[:REGISTERED_ADDRESS]->(:Address)

Importación de Datos

Instalación y Configuración

Instalación de Neo4j

(Específico para archlinux, puede encontrarse el mismo procedimiento para distribuciones basadas en debian en <https://neo4j.com/docs/operations-manual/current/installation/linux/>)

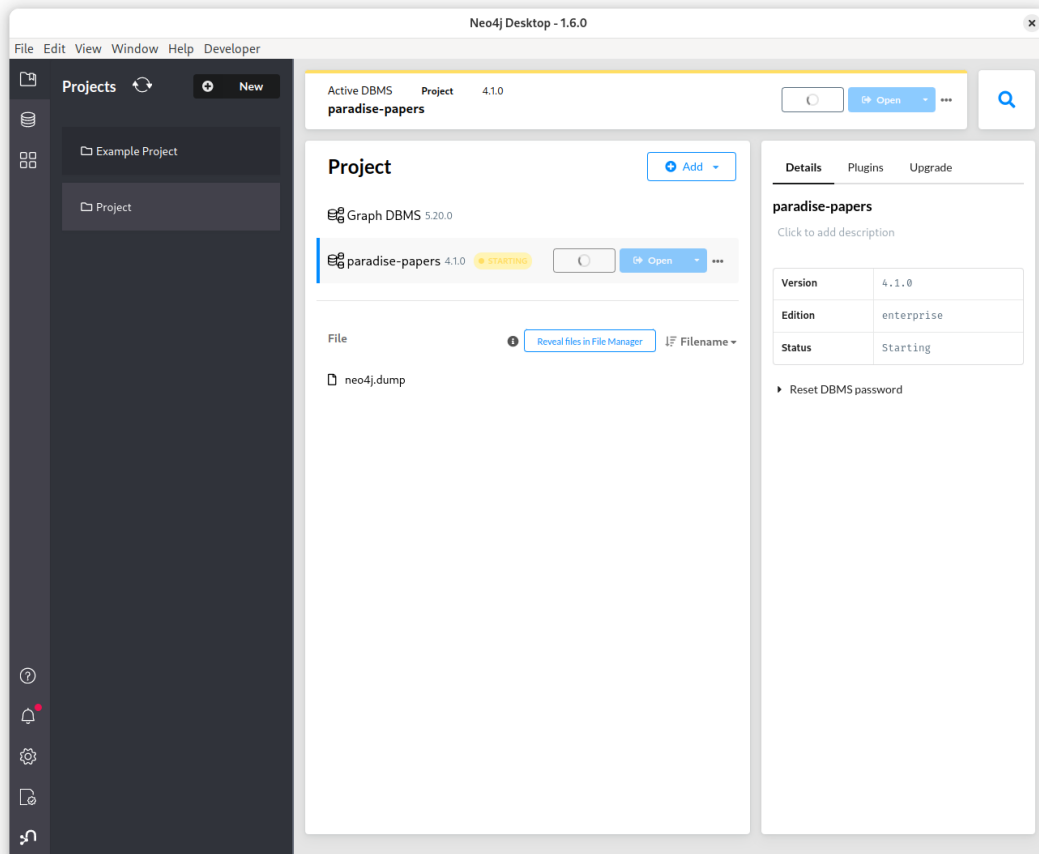
Configuración de versión de java e instalación de Neo4j:

```
sudo archlinux-java set java-23-openjdk
```

```
yay -S neo4j-community
```

Carga de Datos en Neo4j

Consiste únicamente en colocar los archivos .dump en la sección files del proyecto, en Neo4j desktop



Posteriormente podemos explorar la base de datos en cualquiera de las utilidades de la interfaz

neo4j@bolt://localhost:7688/paradise-papers - Neo4j Browser

File Edit View Window Help Developer

paradise-papers\$

paradise-papers\$ MATCH (a:Officer),(b:Officer) WHERE a.name CONTAINS 'Ro...

Overview

Node labels

- * (23)
- Officer (4)
- Entity (15)
- Address (3)
- Other (1)

Relationship types

- * (47)
- OFFICER_OF (18)
- REGISTERED_ADDRESS (21)
- INTERMEDIARY_OF (8)

Displaying 23 nodes, 299 relationships.

Use Ctrl or Shift + scroll to zoom
Don't show again

\$:play <https://guides.neo4j.com/sandbox/icij-paradise-papers/index.html>

use for this purpose than a SQL database or other types of NoSQL databases.

For example, let's say we want to discover the shortest paths between two entity officers through a set of **Entity** or **Address** nodes. This is quite easy with Cypher, Neo4j's graph query language.

```

MATCH (a:Officer),
      (b:Officer)
WHERE a.name CONTAINS 'Ross, Jr'
AND b.name CONTAINS 'Grant'

```

The resulting graph allows us to explore how these people are connected:

Referencias

<https://youtu.be/o4XeBz1hYJQ>

<https://neo4j.com/docs/getting-started/appendix/example-data/>