Introducción a las Bases de Datos de Grafos: Experiencias en Neo4j

Lic. Cristina Vera - Lic. Silvina Migani

Departamento de Informática / Facultad de Ciencias Exactas, Físicas y Naturales / UNSJ
Av. Ignacio de la Roza 590 (O)
Teléfonos: 4260353 - 4260355
civera2@yahoo.com.ar; silvina.migani@gmail.com

Resumen
Las bases de datos NoSQL surgieron como una alternativa de solución a problemas no resueltos eficientemente por las tradicionales bases de datos relacionales. Sin embargo, el término NoSQL abarca diferentes tipos de bases de datos, cada uno con sus características intrínsecas que le confieren un comportamiento más o menos apropiados para diferentes escenarios de aplicación. Las bases de datos de grafos se encuentran dentro de esa gran y variada familia de bases de datos.
Esta línea de investigación surge como una iniciativa de las dos docentes que constituyen el equipo de cátedra de las asignaturas de bases de datos del Departamento de Informática, con la finalidad de profundizar en el estudio de esta nueva y prometedora generación de bases de datos.

Palabras Clave: Bases de Datos – Bases de Datos NoSQL - Bases de Datos de Grafos

Contexto
Las cátedras Bases de Datos y Tópicos Avanzados de Bases de Datos de Departamento

de Informática de la FCEFyN de la UNSJ constituyen el contexto dentro del cual se desarrollan las actividades de investigación.

1. Introducción

El uso masivo de Internet generó nuevas formas de producir y compartir información. La Web 2.0 es una web dinámica y participativa, donde los usuarios son protagonistas activos, generando y compartiendo contenidos, opinando y participando. Esta situación provoca la generación de una enorme cantidad de datos altamente relacionada, que necesita ser almacenada y manipulada eficientemente, con un elevado grado de disponibilidad y que además, comúnmente no se ajusta a una estructura rígida. En este escenario, surgieron las bases de datos NoSQL, ya que las bases de datos relacionales no pudieron satisfacer adecuadamente esas exigencias. Las bases de datos de grafos constituyen una de las alternativas dentro de la gran y variada familia de bases de datos NoSQL. Así, esta propuesta pretende profundizar en el estudio de bases de datos y de gestores de bases de datos basados en grafos (SGBDGs), con el propósito de identificar, distinguir, experimentar y valorar sus características específicas.
Este artículo describe brevemente los temas abordados hasta el momento dentro de la temática planteada:

Sistemas de Bases de Datos basados en Grafos (SBDGs)

Antes de caracterizar a los SBDGs, se presenta la definición matemática de grafo. Un grafo simple *G* está definido por *V(G),* un conjunto finito y no vacío de elementos llamados vértices o nodos, y por *E(G),* un conjunto finito de pares no ordenados de elementos distintos de *V(G)* llamados arcos o aristas. Una arista *{v, w}* une los vértices *v* y *w*, comúnmente denotada como *vw* [1].

Un SGBDG, es un sistema de base de datos específicamente diseñado para poseer las siguientes capacidades [2]:

- Administrar datos de tipo grafo. Es decir, su modelo de datos lógico está basado en alguna de las variantes de la definición matemática básica de grafo, como por ejemplo, grafos dirigidos o no dirigidos, con vértices y arcos etiquetados o no etiquetados, con propiedades en nodos y arcos, hipergrafos e hipernodos [3][4]. Consecuentemente, las operaciones CRUD (Create-Read-Update-Delete, operaciones básicas de los SGBDs) trabajan sobre grafos y sus elementos.

  Sin embargo esto no significa que en el almacenamiento subyacente efectivamente se encuentren grafos. Algunos SGBDGs utilizan almacenamiento nativo [5]; es decir, sus estructuras de datos físicas están diseñadas y optimizadas para almacenar y administrar grafos. También existen gestores que mapean los grafos a otras estructuras [5], como por ejemplo tablas u objetos. En cuanto a la implementación de las conexiones entre nodos, también se presentan diferentes enfoques, algunos que eligen la adyacencia sin índices, y otros que no.
- Satisfacer los principios básicos de todo SGBD, es decir, el almacenamiento persistente, la independencia física/lógica, la integridad y la consistencia de los datos.

Modelo de Grafos vs Modelo Relacional

El modelo de datos relacional y los SGBDs relacionales son ampliamente conocidos y populares. Por ello, a continuación, en la Tabla 1, se presentan algunos de los elementos básicos del enfoque relacional y sus análogos en el modelo de grafos. Cabe aclarar que dicha correspondencia no es absoluta ya que cada pareja de términos presenta particularidades significativas.

Bases de Datos Relacionales	Bases de Datos de Grafos
Filas	Nodos
Columnas	Propiedades
Nombre de las Tablas	Etiquetas en Nodos/Aristas
Claves Foráneas	Aristas entre Nodos

Tabla 1:
Comparación Modelo Relacional – Modelo de Grafos

Algunos SGBDGs presentes en el mercado

Hoy en día existen numerosos SGBDs basados en grafos. A continuación se mencionan en orden de popularidad según [6] algunos: Neo4j, Microsoft Azure Cosmos DB, OrientDB, ArangoDB, Virtuoso y JanusGraph.

Ensayos Realizados

Para poder experimentar las características inherentes a este tipo de sistemas de bases de datos, se siguieron los siguientes pasos:

1. Definición de un escenario sencillo apropiado para ser resuelto con este tipo de tecnología. Así, se concibió un sistema recomendador que brindara información relativa a las sucursales de San Juan Servicio cuando una persona tiene que efectuar un pago, en base a la opinión de sus amigos. Para ello, se registraron las experiencias de pago, ponderadas a través de una calificación (valor numérico entero entre 1 y 10).

2. Elección de un SGBDG donde concretar las experiencias. Se decidió usar Neo4j, sigue el modelo de grafo de propiedades y es una herramienta muy útil para prototipado rápido y para hacer pruebas de concepto, ya que es sencilla y no presenta complicaciones en la configuración [7]. Además, provee el lenguaje Cypher que permite expresar las operaciones CRUD de manera bastante natural [8].

3. Implementación del ejemplo planteado. Luego de instalar Neo4j Desktop 1.1.15 se generó el grafo correspondiente a través de operaciones escritas en el lenguaje Cypher.

Algunas de las operaciones de creación de nodos y aristas se muestran a continuación:

CREATE (Maria:Persona {nombre:'María', fechanac:'26-01-1990'})

CREATE (Suc1:Sucursal {nombre:'Casa Matriz', localidad:'San Juan', Direccion: 'Scalabrini Ortiz 1285 N', horaten:'8-13'})

CREATE (Maria)-[:Paga {califvisita:6}] ->(Suc1)

CREATE (Maria)-[:Amigo]->(Juan)

Analizando la relación Paga, se observa que contiene la propiedad calificación. En el ejemplo, María calificó a la Casa Matriz (Suc1) con 6.

El grafo resultante luego de la ejecución de las sentencias de creación de todos los nodos y aristas se muestra a continuación en la Figura 1.

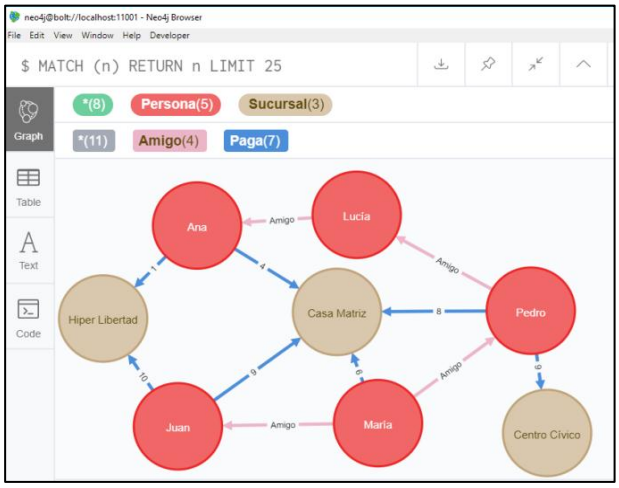

Figura 1: Grafo ejemplo

Asimismo, a modo de ejemplo, se muestran a continuación algunas de las consultas implementadas para el sistema recomendador propuesto.

C1: Obtener para cada sucursal, su calificación promedio.

MATCH (S:Sucursal)-[Pa:Paga]-() WITH S, AVG(Pa.califvisita) AS prom
return S.nombre,prom

La Figura 2 muestra el resultado correspondiente[1].

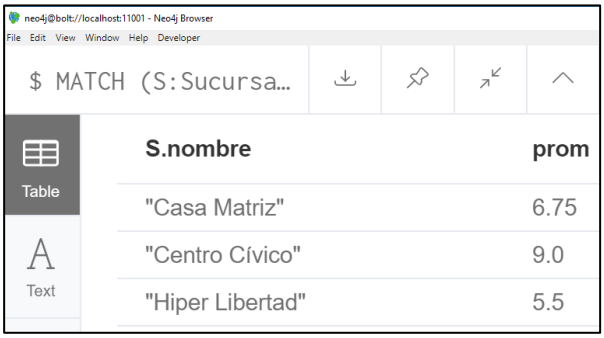

Figura 2: Resultado consulta C1

[1] Esta consulta no devuelve un grafo, por eso el resultado se muestra como una tabla.

C2: Obtener la sucursal con máxima calificación promedio.

MATCH (s:Sucursal)-[Pa:Paga]-() WITH s, AVG(Pa.califvisita) AS prom
RETURN s ORDER BY prom DESC
LIMIT 1

El resultado obtenido se presenta en la Figura 3.

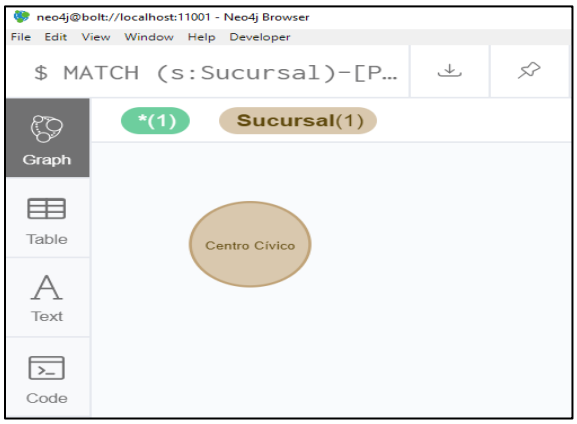

Figura 3: Resultado de la consulta C2

C3: Considerando las calificaciones de los amigos de Juan, obtener la sucursal con mayor calificación promedio.

MATCH (p:Persona{nombre:'Juan'})-[A:Amigo]-(p2:Persona)-[pa:Paga]-> (s:Sucursal) WITH s, AVG(pa.califvisita) AS prom
RETURN s as Sucursal,prom as Promedio
ORDER BY prom DESC
LIMIT 1

En la Figura 4 se puede observar el resultado.

Figura 4: Resultado de la consulta C3

C4: Considerando las calificaciones de los amigos de Juan y de los amigos de los amigos de Juan, obtener la sucursal con mayor calificación promedio.

*MATCH (p:Persona{nombre:'Juan'})-[A:Amigo*1..2]-(p2:Persona)-[pa:Paga]->*
(s:Sucursal) WITH s, AVG(pa.califvisita) AS prom
RETURN s as Sucursal,prom as Promedio
ORDER BY prom DESC
LIMIT 1

El resultado obtenido se presenta en la Figura 5.

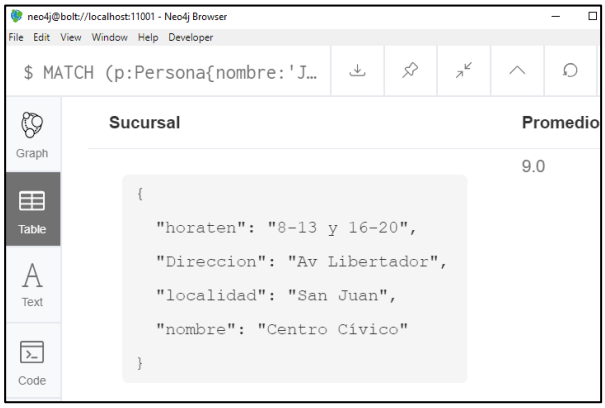

Figura 5: Ejecución de la consulta C4

Como se puede observar, la sucursal recomendada a Juan al considerar los pagos realizados por sus amigos (C3) no es la misma que la obtenida al considerar además de los pagos de los amigos de Juan, los pagos de los amigos de los amigos de Juan (C4); ya que se consideran instancias diferentes de la relación Paga.

En base a los ejemplos planteados e implementados, se pudo distinguir y valorar el potencial de esta tecnología, fundamentalmente en lo que concierne a:

- La gran flexibilidad en la estructura de los datos. Tanto los nodos como las aristas no necesitan ajustarse a una estructura predeterminada.

- El hecho de contar con un lenguaje de consulta declarativo potente e intuitivo, que hace muy sencilla la tarea de codificar consultas.
- La posibilidad de navegar eficientemente a través de relaciones complejas, incluso recursivas (relaciones entre nodos del mismo "tipo") a cualquier nivel de profundidad, incluso no fijo. Capacidad no ofrecida convenientemente por los sistemas SQL.

Líneas de Investigación y desarrollo

La investigación tiene como eje central el estudio y la experimentación de bases de datos NoSQL basadas en grafos.

Resultados y Objetivos

Los propósitos fundamentales establecidos fueron:

- Profundizar en el estudio de las bases de datos de grafos en el marco de los SGBD NoSQL.
- Investigar y experimentar SGBDs dentro de esta tipología.
- Asesorar trabajos finales en el área.
- Elaborar material bibliográfico.
- Fortalecer las cátedras de Bases de Datos del Departamento de Informática.

Sin embargo, dado que el trabajo de investigación se encuentra en sus inicios, los resultados obtenidos hasta el momento son:

- Se han estudiado las características intrínsecas del modelo de datos de grafos, identificando clasificaciones dentro de este grupo.
- Se han investigado gestores de bases de datos de grafos presentes en el mercado.
- Se ha profundizado en el estudio de Neo4j Desktop 1.1.15 como SGBD de experimentación.
- Se han definido e implementado algunos ejemplos concretos en Neo4j.
- Se está desarrollando una tesis de grado y una de maestría sobre la temática NoSQL, que incluyen los SGBDGs.

Formación de Recursos Humanos

- Se está desarrollando una tesis de grado y una de maestría sobre la temática NoSQL, que incluyen los SGBDGs.

Bibliografía

[1] Wilson, R. J., Wilson, R. J., Wilson, R. J., & Wilson, R. J. (1972). Introduction to graph theory (Vol. 107). London: Longman.

[2] Angles, R., & Gutierrez, C. (2018). An introduction to Graph Data Management. In Graph Data Management (pp. 1-32). Springer, Cham.

[3] Sasaki B., Chao J. & Howard R. (2018). Graph Databases for Beginners.

[4] Introduction to Graph Databases.

Disponible en https://neo4j.com/graphacademy/online-training/introduction-to-neo4j/part-1/. Fecha última visita Marzo 2019.

[5] DB-Engines Ranking of Graph DBMS. Disponible en https://db-engines.com/en/ranking/graph+dbms. Fecha última visita Marzo 2019.

[6] Robinson, I., Webber, J., & Eifrem, E. (2013). Graph databases. " O'Reilly Media, Inc.".

[7] Sitio oficial de Neo4j. Disponible en https://neo4j.com Fecha última visita Marzo 2019.

[8] The Neo4j Cypher Manual v3.5. Disponible en https://neo4j.com/docs/cypher-manual/3.5/. Fecha última visita Marzo 2019.