Base de datos basado en grafos para un análisis de diversas perspectivas

Rodrigo Alexander Mamani Sucacahua

Orientador: Prof Dr./Mag./Ing. Nombre del Asesor

Plan de Tesis presentado la Escuela Profesional Ciencia de la Computación como paso previo a la elaboración de la Tesis Profesional.

UNSA - Universidad Nacional de San Agustín de Arequipa
Mayo de 2022

Abreviaturas

Índice

Índice de cuadros

Índice de figuras

1. Motivación y Contexto

Hoy en dia, las universidades públicas y privadas presentan la necesidad de contactar,contratar y encontrar investigadores en áreas en específico, ya sea para ayudar a los alumnos en su asesoramiento para sus proyectos de tesis o para dirigir proyectos de pregrado, para estos fines se evaluan de diversas maneras y puntos de referencia dados por los interesados. Esta toma de decisiones ayuda a los interesados tener veracidad y respaldo de calidad en el área de investigación.Para esto se opta a buscar en bases de datos públicas las cuales brindan dicha información realizando largas búsquedas para encontrar a los investigadores de su interés. Concretamente en el Perú la página de DINA de Concytec brinda información de los investigadores como bibliografía, experiencia profesional, datos académicos, producción científicas, proyectos de investigación entre otros dotas los cuales pueden ser de gran utilidad para dicho objetivo .

2. Definición del Problema

Las búsquedas largas en páginas como DINA son muy engorrosas ya que no solo es encontrar a un sector de investigadores de su interés sino también leer los datos en forma de texto y sus curriculum vitae respectivos. Además, existen datos abstractos del interés académico que no se muestran en dicha página las cuales las relaciones de estas a menudo son hasta necesario consultar con el mismo investigador para saber si tiene lo que se busca o no, como que si tiene contacto con el extrangero, si conoce a colegas o investigadores de su rama entre otras cosas. Al plantear el traslado de los datos a una base de datos lo mas común es hacerlo a una base de datos relacional en lenguaje SQL , sin embargo, al momento de hacer consultas para encontrar un dato determinado se usan demasiados comandos de consultas JOIN , los cuales, si bien es cierto que es mas fácil la interacción con datos existen bases de datos que pueden brindar los datos requeridos de manera óptima como una base de datos basada en grafos.

3. Objetivos

Los objetivos se relacionan directamente con el problema definido, ya que el objetivo es la forma en la que pensamos resolver el problema propuesto.

Una vez que queda definido cuál es el objetivo principal, se deben desprender los objetivos específicos según la estrategia, metodología y técnicas que se piensan usar para resolver el problema.

3.1. Objetivo General

Análisis de diversas perspectivas de una base de datos basada en grafos

3.2. Objetivos Específicos

- Creacción de Base de datos basada en grafos usando datos de uso libre.
- Uso el lenguaje Cypher y una Herramienta para construcción de base de datos basada en grafos.
- Realizar un análisis bajo algunos parámetros concisos.
- Mostrar grafos representativos de la base de datos basado en grafos.

4. Trabajos Relacionados

El enfoque dado para el muestreo de investigadores es eficiente para la búsqueda concreta de los investigadores de interés y se demuestra el funcionamiento de esta base de datos basa den grafos y cuan eficiente es este modelo [Afandi and Wahyuni, 2020]. Dina no es la única fuente de datos de investigadores a nivel mundial, en este artículo plantea la visualización de datos de DSPACE en Neo4j y Dephi herramientas de visualización de código de acceso libre y apliaciones desktop [Aryani et al., 2017]. Para tener un buen análisis de una herramienta de creación de base de datos basado en grafos como Neo4j se debe evaluar como el back-end trabajado en esta aplicaciónayuda a los desarrolladores para su fácil entendimiento[Holzschuher and Peinl, 2013].
Para analizar a fondo la creación de la base de datos se deben evaluar la relación entre atributos y la asociación que estos presentan con la base de datos.[Lu et al., 2017].También se deben evaluar las consultas y transacciones que se pueden realizar en la base de datos creada en Neo4j[Vukotic et al., 2015].
Neo4j al ser una herramienta basada en grafos, es evidente que la base de datos creada sea basada en grafos, esta puede ser planteada en una base de datos distribuida como las social networks la cual ,tal como la base de datos general, de Neo4j esta también puede ser visualizada de manera óptima[Agudo Merino, 2020].
Como toda aplicación Neo4j tiene ventajas y desventajas en su uso, las cuales mediante comparaciones con diversas aplicaciones pueden ayudar al usuario a decidir si es factible o no usar dicha herramienta[Fernandes and Bernardino, 2018].Las limitaciones que esta herramienta presenta pueden ser evaluados por su interacción de la base de datos con las consultas y con los datos los cuales demuestran que Neo4j es una herramienta optimizable[Miller, 2013].
El análisis de una herramienta como Neo4j también puede ser evaluado por el tiempo de demora de la creación de la base de datos y su complejidad de consultas comparadas con otra herramienta como PostgreSQL.[Stothers and Nguyen, 2020]. Neo4j no

solo sirve para bases de datos convensionales , sino también ayudan para el mejor control de la semántica e interacción de las bases de datos como las redes sociales y web semántica[Guia et al., 2017].

5. Propuesta

La propuesta consta de un Pipeline (un gráfico) que describe el proceso planteado para resolver el problema. Si aún no se sabe exactamente que técnicas se van a utilizar por lo menos se debe poner el grupo de técnicas posibles para cada una de las etapas o el area de conocimiento.

Ejemplo: *En la Figura 1 se puede observar un ejemplo de pipeline.*

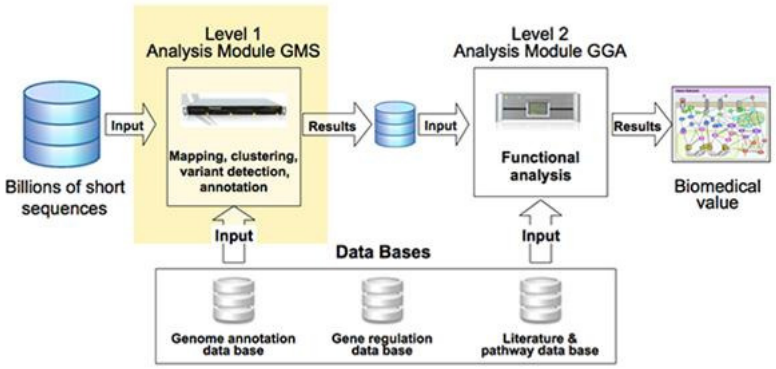

Figura 1: Pipeline del modelo propuesto.

6. Cronograma de Actividades

Aquí se debe describir las actividades que se llevarán a cabo y un aproximado del tiempo que tomará cada una de ellas (dia/mes de inicio y fin). Este cronograma deberá se definido junto al asesor y deberá contener el detalle necesario para explicar el flujo de trabajo que realizará el alumno.

Ejemplo: *En el Cuadro 1 se puede observar un ejemplo de un Cronograma de Actividades.*

7. Indice Tentivo de la Tesis

Aquí se debe incluir un posible índice para la tesis una vez que ésta esté terminada. Posiblemente este índice se cambiará cuando se escriba la tesis final, pero por lo menos

Cuadro 1: Posible Cronograma de Actividades para el desarrollo de la Tesis.

Actividad	Inicio Aprox.	Fin Aprox.
Elaboración del Proyecto de Tesis	*25-jul-13*	*30-jul-13*
Presentación y Aprobación del Proyecto	*01-ago-13*	*08-ago-13*
Redacción de la Parte Teórica de la Tesis	*09-ago-13*	*09-set-13*
Implementación de las técnicas a ser usadas	*10-set-13*	*25-oct-13*
Realización de las pruebas y escritura de resultados	*26-oct-13*	*26-nov-13*
Presentación del Borrador de Tesis	*27-nov-13*	*04-dic-13*
Correción de observaciones al Borrador de Tesis	*05-dic-13*	*12-dic-13*
Presentación de la Versión Final de la Tesis	*22-dic-13*	*24-dic-13*
Sustentación de la Tesis	*26-dic-13*	*28-dic-13*

debería ser un bosquejo para tener una idea de los contenidos que se piensan tratar y en que orden.

Ejemplo: *La tesis a ser desarrollada como producto de este Plan de Tesis tendra tentantivamente el siguiente Indice de Contenidos:*

1. *Resumen.*
2. *Abstract.*
3. *Introducción.*
   a) *Motivación y Contexto.*
   b) *Definición del Problema.*
   c) *Objetivos.*
      1) *Objetivo Principal.*
      2) *Objetivos Específicos.*
4. *Conceptos sobre Bases de Datos.*
   a) *Bases de Datos Relacionales.*
   b) *PostgreSQL.*
   c) *Estructuras de Datos implementadas en PostgreSQL.*
5. *Trabajos Relacionados.*
   a) *Otros motores de Bases de Datos.*
   b) *Optimizaciones previas a PostgreSQL.*
   c) *Optimizaciones a otros motores de Bases de Datos.*
6. *Propuesta.*
   a) *Mejoras propuestas al algoritmo de Inserción.*

   b) *Mejoras propuestas al algoritmo de Actualización.*
   c) *Mejoras propuestas al algoritmo de Eliminación.*
   d) *Comparación con mejoras previas.*
   e) *Algoritmo General Propuesto.*
   f) *Pipeline propuesto.*
7. *Experimentos y Resultados.*
   a) *Descripción de la metodologñia utilizada para los expermientos.*
   b) *Arquitectura utilizada.*
   c) *Descripcion de los experimentos realizados.*
   d) *Resultados.*
      1) *Comparacion con Propuesta A (de la literatura revisada).*
      2) *Comparacion con Propuesta B (de la literatura revisada).*
8. *Conclusiones.*
9. *Recomendaciones.*
10. *Trabajos Futuros.*

Referencias

[Afandi and Wahyuni, 2020] Afandi, M. I. and Wahyuni, E. D. (2020). University research graph database for efficient multi-perspective data analysis using neo4j. In *2020 6th Information Technology International Seminar (ITIS)*, pages 286–290. IEEE.

[Agudo Merino, 2020] Agudo Merino, A. (2020). Base de datos distribuida de grafos con neo4j para el record linkage de redes sociales.

[Aryani et al., 2017] Aryani, A., Wang, J., Zhang, H., Xiang, A., Zhou, Z., and Wang, K. (2017). Visualising research graph using neo4j and gephi.

[Fernandes and Bernardino, 2018] Fernandes, D. and Bernardino, J. (2018). Graph databases comparison: Allegrograph, arangodb, infinitegraph, neo4j, and orientdb. In *Data*, pages 373–380.

[Guia et al., 2017] Guia, J., Soares, V. G., and Bernardino, J. (2017). Graph databases: Neo4j analysis. In *ICEIS (1)*, pages 351–356.

[Holzschuher and Peinl, 2013] Holzschuher, F. and Peinl, R. (2013). Performance of graph query languages: comparison of cypher, gremlin and native access in neo4j. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, pages 195–204.

[Lu et al., 2017] Lu, H., Hong, Z., and Shi, M. (2017). Analysis of film data based on neo4j. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pages 675–677. IEEE.

[Miller, 2013] Miller, J. J. (2013). Graph database applications and concepts with neo4j. In *Proceedings of the southern association for information systems conference, Atlanta, GA, USA*, volume 2324.

[Stothers and Nguyen, 2020] Stothers, J. A. and Nguyen, A. (2020). Can neo4j replace postgresql in healthcare? *AMIA Summits on Translational Science Proceedings*, 2020:646.

[Vukotic et al., 2015] Vukotic, A., Watt, N., Abedrabbo, T., Fox, D., and Partner, J. (2015). *Neo4j in action*, volume 22. Manning Shelter Island.