



Base de datos basado en grafos de investigadores de Concytec 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

1. Motivación y Contexto	6
2. Definición del Problema	6
3. Objetivos	7
3.1. Objetivo General	7
3.2. Objetivos Específicos	7
4. Trabajos Relacionados	7
5. Propuesta	8
6. Cronograma de Actividades	8
7. Índice Tentativo de la Tesis	8

Índice de cuadros

1. Posible Cronograma de Actividades para el desarrollo de la Tesis. 9

Índice de figuras

1.	Pipeline del modelo propuesto.	8
----	--	---

1. Motivación y Contexto

Listado de artículos de investigación ordenados según su relevancia para esta investigación.

- University Research Graph Database For Efficient Multi-Perspective Data Analysis Using Neo4j [Afandi and Wahyuni, 2020].
- Graph Databases Comparison: AllegroGraph, ArangoDB, InfiniteGraph, Neo4J, and OrientDB. [Fernandes and Bernardino, 2018].
- Base de datos distribuida de grafos con Neo4j para el Record Linkage de Redes Sociales [Agudo Merino, 2020].
- Performance of graph query languages: comparison of cypher, gremlin and native access in neo4j [Holzschuher and Peinl, 2013]
- Can Neo4j Replace PostgreSQL in Healthcare? [Stothers and Nguyen, 2020].

Hoy en día, 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 evalúan 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 científica, proyectos de investigación entre otros datos 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 extranjero, 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 más 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 más 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

***Ejemplo:** Mejorar las estructuras de datos implementadas en el motor de base de datos PostgreSQL para trabajar eficientemente con datos geo-referenciados en tiempo real.*

3.2. Objetivos Específicos

Ejemplo:

- *Estudiar las estructuras de datos del motor de bases de datos PostgreSQL.*
- *Mejorar los algoritmos de insercion de la estructuas de datos de PostgreSQL.*
- *Mejorar los algoritmos de actualizacion de la estructuas de datos de PostgreSQL.*
- *Mejorar los algoritmos de eliminacion de la estructuas de datos de PostgreSQL.*
- *Realizar pruebas de desempeño entre los algoritmos antiguos y algoritmos nuevos.*

4. Trabajos Relacionados

En esta sección se desarrolla el problema más profundamente y se habla sobre los trabajos relacionados al mismo.

Se debe escribir todo lo necesario para que el lector pueda entender profundamente el problema que se pretende resolver. Aquí es importante referenciar los trabajos clásicos más resaltantes, de los cuales se extraen los conceptos principales.

Una vez que se describe el problema, se deben mencionar los trabajos relacionados al mismo. Aquí debemos destacar los trabajos donde se proponen soluciones alternativas a nuestro problema. Por cada trabajo revisado se debe incluir un párrafo donde se resume el enfoque propuesto en el trabajo, así como la forma en la que puede servirnos y además cuáles son sus deficiencias.

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.

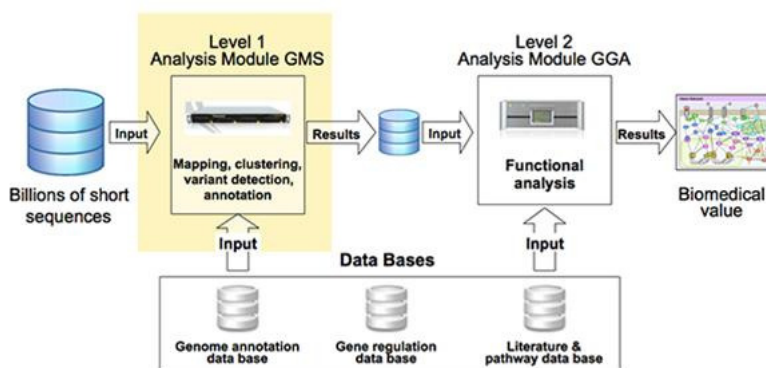


Figura 1: Pipeline del modelo propuesto.

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 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 tentativamente el siguiente Indice de Contenidos:

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

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

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 metodología utilizada para los experimentos.*
- b) *Arquitectura utilizada.*
- c) *Descripción de los experimentos realizados.*
- d) *Resultados.*
 - 1) *Comparación con Propuesta A (de la literatura revisada).*
 - 2) *Comparación 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.
- [Fernandes and Bernardino, 2018] Fernandes, D. and Bernardino, J. (2018). Graph databases comparison: Allegrograph, arangodb, infinitegraph, neo4j, and orientdb. In *Data*, pages 373–380.
- [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.
- [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.