2020 Information Technology International Seminar
Surabaya, Indonesia, October 14-16 2020

University Research Graph Database For Efficient Multi-Perspective Data Analysis Using Neo4j

Mohamad Irwan Afandi
Information Systems Department
Universitas Pembangunan Nasional Veteran Jawa Timur
Surabaya, Indonesia
mohamadafandi.si@upnjatim.ac.id

Eka Dyar Wahyuni
Information Systems Department
Universitas Pembangunan Nasional Veteran Jawa Timur
Surabaya, Indonesia
ekawahyuni.si@upnjatim.ac.id

2020 6th Information Technology International Seminar (ITIS) | 978-1-7281-7726-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ITIS50118.2020.9320965

Abstract— **In general, research-related data are modeled using a relational database optimized for transaction processing. In many cases, this solution is effective and efficient enough to answer basic queries and simple reporting requirements. However, when users request a more-in-depth, more expansive, multi-perspective, and sometimes more abstract analysis, the relational database struggles to provide answers. This study proposes a research graph database implemented using neo4j as an effort to answer the problems. The database consists of a core model and an extension model. The core model represents scientific articles-related data loaded with real data scraped from Google Scholar. The extension model indicates research and community engagement activities done by researchers loaded manually. The database enables the university to analyze researchers' individual and collaborative performances with fellow researchers inside and outside universities. The study concludes that the research graph database implementation is more efficient in answering similar questions than the relational database implementation.**

Keywords— graph database, research data, efficient data analysis, multi-perspective, neo4j

I. Introduction

Organizations, including universities, face challenges concerning realizing data integration. Typical problems found in universities are fragmented data, flawed and incomplete data model design, managed separately by different systems across the organization. It is often further worsened with no documentation. Reports generation is also a long, painful process. Simultaneously, the analysis activity (e.g., SWOT Analysis) is limited in depth and breadth, mostly caused by the lack of data integrity and connectivity.

In line with the continuing increase of demands from the government to universities for higher quality, especially in the research and community engagement area where new quality measures are in effect, any university needs to build better solutions to answer those requirements. It must look carefully at their current systems and data models and develop new ideas to keep up with the dynamic environment and answer future organizations challenging questions.

In recent years graph database has been increasingly getting attention and researched extensively for many purposes such as (i) analysis of film and analysis of lightning and transmission failure rate relationship [1][2]; (ii) recommendation on e-commerce, recipe, book, movie, and retail [3][4]–[7]; (iii) information retrieval and discovery [8][9]; (iv) data integration [10]. It also gains popularity in the industry with several prominent organizations, including NASA (National Aerospace Service Agency), known for using the approach [11].

This research aims to produce a graph of researchers and their publications from where knowledge workers can query against to analyze the network of authors, organizations, keywords, research interests, and others. The graph resulted from this study is hoped to be easily updated, extended, and up-scaled by inserting more nodes and adding relations to domain knowledge to support semantic searches.

II. Methodology

There are many search engines for academic literature available on the Internet today such as Google Scholar, Google Books, Microsoft Academic, Researchgate, Science.gov, Refseek, ERIC (Educational Resources Information Center), WorldWideScience, iSeek, VLRC (Virtual Learning Resources Center), BASE (Bielefeld Academic Search Engine), PubMed, and others. Reports by several studies favor Google Scholar over the others. They present that Google Scholar is widely used globally due to its high recall, free access, and broad coverage [12][13]. Among those initiatives which rely on Google Scholar is a platform named SINTA (Science and Technology Index). Ristekbrin (Research and Technology Ministry / National Research and Innovation Body) of Indonesia currently hosted SINTA. SINTA utilizes data from Google Scholar [14] and other sources and then calculates researchers' capabilities, institutions, and journals in Indonesia. The results are then made available for free to the public. Google Scholar's advantages worth highlighting are the facts that Google Scholar offers more updated academic references. There are also APIs (Application Programming Interfaces) in various programming platforms built by developers that researchers, practitioners, and the general public can use to connect to Google Scholar and retrieve data from it freely for academic and research purposes. Therefore, it fits the requirements of this study.

This research used Google Scholar as the primary source of scientific articles related-data to produce a graph database. The secondary data sources are research activities data retrieved from various sources ranging from university research databases, spreadsheets, and the Internet. Fig. 1 depicts the detail of the steps taken in this study.

978-1-7281-7726-7/20/$31.00 ©2020 IEEE

As an initial step, the study identified the data available in and offered by Google Scholar through direct observation of the https://scholar.google.com and literature study. Next, based on the data of entities and possible relations between them, questions were generated. The initial questions list was never intended to be final nor fixed but dynamic. However, we assessed every question dimension of 4W (*What, Where, When, Who*) and 1H (*How*) to see whether the data collected answered the questions.

The following step was designing the graph data model consisting of nodes, labels, relationships, and property keys based on the previous actions and implementing it on the neo4j platform. The graph queries, written in CYPHER, are intended to answer questions and test the model were then formulated.

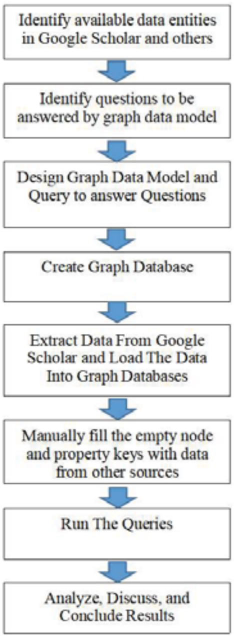

Identify available data entities in Google Scholar and others
Identify questions to be answered by graph data model
Design Graph Data Model and Query to answer Questions
Create Graph Database
Extract Data From Google Scholar and Load The Data Into Graph Databases
Manually fill the empty node and property keys with data from other sources
Run The Queries
Analyze, Discuss, and Conclude Results

Fig. 1. Research Methodology

Once the model was tested and validated using dummy data, the next step was to extract data from Google Scholar using Scholarly, a python API to retrieve google scholar data, and subsequently create the graph and executing it.

Research activity related data were filled manually using data gathered earlier. After the graph database loaded all of the data, it performed the queries to answer questions previously prepared. We then compared the graph queries' performance to SQL's performance run against the relational database storing similar data.

III. Result And Discussion

We conducted this research in a university setting. Hence, we perceived the term 'researcher' in Indonesia's higher education context notable for the Tri Dharma (3 obligations) activities: teaching, researching, and community engagement. Therefore, in Indonesia, a university lecturer is also, by default, a researcher.

The graph database resulted from this case study consists of 12 Node Labels, 15 Relationships, and 33 Property Keys. Table I shows all node label names, and Table II displays the relationship between nodes.

TABLE I. Node Label

No	Node Label	Description
1	Affiliation	It represents a researcher's affiliation, such as a university or a research institution, usually written in a scientific article.
2	Article	It represents scientific articles or reports written by Researchers
3	DbIndexer	It describes indexing database services of scientific articles provided by specific organizations, such as Scopus, Web Of Science.
4	IntelPropRight	It typifies any intellectual property right owned by any Researcher
5	Organization	It represents any organization of any kind which researchers affiliate with, receive projects grant from, and have their intellectual property approved
6	Person	It depicts any person in general who can be a researcher or any other type of role or job
7	Publisher	It portrays a particular type of organization which offers publication products
8	Pubname	It illustrates any publication product offered by a Publisher
9	QualityMeasure	It represents any quality measurement of a publication product offered by an indexing database service either directly offered by the database or indirectly by a third party
10	RegisterOffice	A particular type of organization which provides intellectual property rights services from applying to awarding
11	ResearchProject	Any project which can be either research or community engagement type, self-funded or granted by an individual or a group of organizations
12	Researcher	It represents a select type of Person who works either partially or entirely as a researcher.

TABLE II. Nodes' Relation

No	Node	Relation (Action) Name	Node
1	Researcher	AFFILIATES_WITH	Affiliation
2	Researcher	AUTHORS	Article
3	Researcher	CO_AUTHORS	Article
4	Researcher	CO_LEAD	ResearchProject
5	Researcher	CO_OWNS	Intellectual PropertyRight
6	Organization	GRANTS	ResearchProject
7	RegisterOffice	GRANTS_IPR	Intellectual PropertyRight
8	DbIndexer	HAS_MEASUREMENT	QualityMeasure
9	Article	INDEXED_IN	DbIndexer
10	Researcher	LEAD	ResearchProject
11	Researcher	OWNS	Intelletual PropertyRight
12	ResearchProject	PRODUCES	Intellectual PropertyRight
	ResearchProject	PRODUCES	Article
13	Article	PUBLISHED_IN	PubName
14	PubName	RANKED	QualityMeasure
15	Researcher	REG_AUTH_IN	DbIndexer

TABLE III. EXAMPLE OF MANUAL DATA LOADING

No	Query	Remarks

1	MATCH (a:Researcher), (b:IntelPropRight)
	WHERE a.name STARTS WITH 'Moha' AND a.name ENDS WITH 'Afandi' and b.title CONTAINS 'Mobile Business Intelli'
	CREATE (a)-[c:OWNS]->(b)
	RETURN a,b,c
		Three similar Researcher nodes own the same Intellectual Property Right, so the nodes have to be connected manually

2	MATCH (a:Researcher), (b:DbIndexer)
	WHERE a.name STARTS WITH 'Moha' and a.name ENDS WITH 'Afandi' and b.name = 'Scopus'
	CREATE (a) - [c:REG_AUTH_IN {reg_id: '57193856616', name_of_id: 'Scopus Author ID'}] → (b)
	RETURN a, b
		Scopus ID is not available in Google Scholar, so we had to manually insert it.

We then loaded the graph database with data scraped from Google Scholar. Additionally, since not all data are available in Google Scholar, manual data loading was conducted using the query shown in Table III. For example, even though researcher data are already in the graph database, a property such as Scopus ID is specific to a particular database indexer and not provided in Google Scholar. Therefore, after collecting the Scopus ID from different sources, it has to be inserted manually.

TABLE IV. EXAMPLE OF MULTI-PERSPECTIVE INQUIRIES

Case	Inquiries
1	Find **Organizations** or **Researchers** or any entities who are related to **Research Projects** which **Produce Scientific Articles**
2	Find journals **(PubName)** that Publish Articles has **Jariyah** as the Main **Author** and find **All Researchers** who are her **Co-Authors**.
3	Assess how an **Affiliation/Organization** collaborate with other **Affiliations/Organizations**?

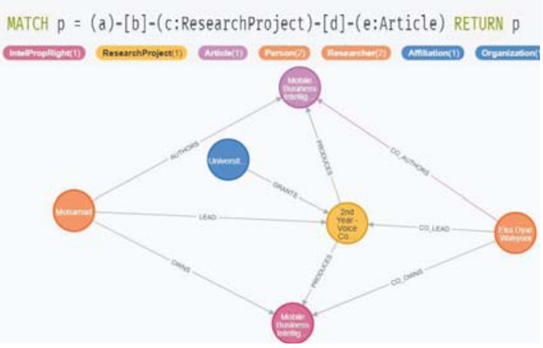
We developed several test scenarios to test the graph database as presented in Table III. The first case expresses information request for any node related to research projects which produce scientific articles. This question may arise when the university executives want to know which projects have scientific papers and give credits or rewards to any individual or organization involved in the projects.

The second scenario may transpire when there is a need to find out how far and wide a particular university researcher has been reaching out beyond his/her self, department, university, and discipline to collaborate with other researchers and produce scientific articles. The more researchers who collaborate and the farther and broader the networks that researchers build, the better the universities' position will be in local, national, regional, and global engagement.

The last case represents a call for a view on how a particular university collaborates with several organizations or affiliations through its researchers. The answer to this inquiry is essential to see which organizations have stronger ties compared to the others.

A. Query Execution For Case 1

Fig. 2. provides an answer for the first test scenario.

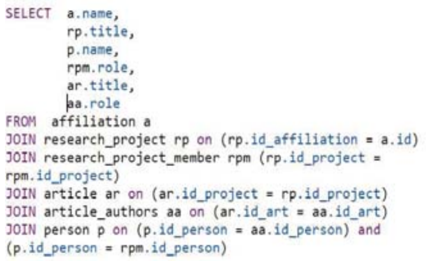


```
MATCH p = (a)-[b]-(c:ResearchProject)-[d]-(e:Article) RETURN p
```

Fig. 2. Cypher query and result for case 1

The graph query written in Cypher is efficient and straightforward. It arguably can be formulated and done by less experienced knowledge workers given the prior training and equipped with information such as node and node relation. Moreover, the visual information displayed is also informative and suitable for different viewers' level, from knowledge workers to executives.

A longer query to answer the same question in a relational database would be similar to Fig. 3. The difference is clear. In a relational database, the requirement has to be very explicit upfront to be correctly and carefully translated into query before deciding which tables to join on which columns and which fields need to be retrieved. Indeed, it requires some level of technical expertise to execute it.



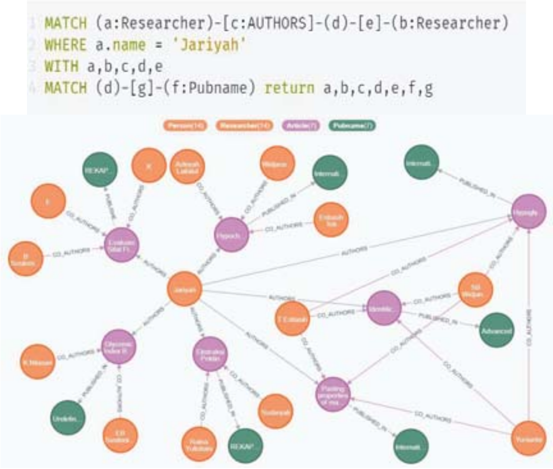
```
SELECT  a.name,
        rp.title,
        p.name,
        rpm.role,
        ar.title,
        aa.role
FROM  affiliation a
JOIN research_project rp on (rp.id_affiliation = a.id)
JOIN research_project_member rpm (rp.id_project =
rpm.id_project)
JOIN article ar on (ar.id_project = rp.id_project)
JOIN article_authors aa on (ar.id_art = aa.id_art)
JOIN person p on (p.id_person = aa.id_person) and
(p.id_person = rpm.id_person)
```

Fig. 3. Corresponding SQL query to answer case 1

B. Query Execution For Case 2

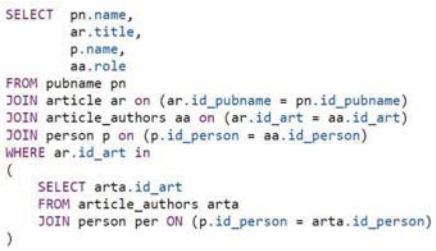
The Cypher query in Fig. 4. is abstract and more natural since the underlying relations and nodes are not clearly defined but only assumed to exist. However, the returned graphs are informative and quickly give useful insights, which relational database is difficult to provide.

Using a relational database as depicted in Fig. 5., the query requires longer steps, joins, and time to execute it, especially when the database size is large. Like the previous query, it is too technical and not easily understood by general people and even knowledge workers.



```
MATCH (a:Researcher)-[c:AUTHORS]-(d)-[e]-(b:Researcher)
WHERE a.name = 'Jariyah'
WITH a,b,c,d,e
MATCH (d)-[g]-(f:Pubname) return a,b,c,d,e,f,g
```

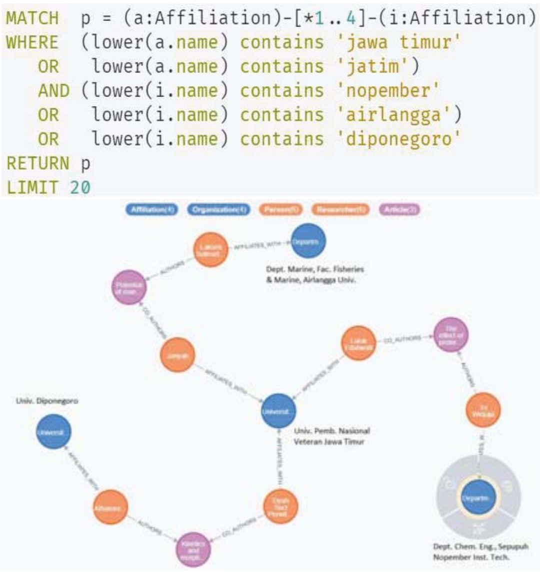
Fig. 4. Cypher query and result for case 2



```
SELECT  pn.name,
        ar.title,
        p.name,
        aa.role
FROM pubname pn
JOIN article ar on (ar.id_pubname = pn.id_pubname)
JOIN article_authors aa on (ar.id_art = aa.id_art)
JOIN person p on (p.id_person = aa.id_person)
WHERE ar.id_art in
(
    SELECT arta.id_art
    FROM article_authors arta
    JOIN person per ON (p.id_person = arta.id_person)
)
```

Fig. 5. Corresponding SQL query for case 2

C. Query Execution For Case 3



```
MATCH  p = (a:Affiliation)-[*1..4]-(i:Affiliation)
WHERE  (lower(a.name) contains 'jawa timur'
   OR   lower(a.name) contains 'jatim')
   AND (lower(i.name) contains 'nopember'
   OR   lower(i.name) contains 'airlangga')
   OR   lower(i.name) contains 'diponegoro'
RETURN p
LIMIT 20
```

Fig. 6. CYPHER query and results for Case 3

For case 3, the question expects relation(s) exist(s) between organizations in one way or another, whether directly or indirectly. A direct relation (one link away) would be something like *Organization X HAS_MOU_WITH Organization Y*. On the other hand, an indirect connection would likely manifest into something like *Organization X EMPLOYS a Person P who WRITES an Article which is CO_WRITE by a Person P2 who WORKS for Organization Y*. In this case, there are four relations in between Organization X and Organization Y. The answer in CYPHER is as natural as it can be, as depicted in Fig. 6.

As demonstrated in case 1, case 2, and case 3, the graph database and query do the jobs of answering inquiries efficiently compared to the relational database. These simple experiments are in line with previous studies that compared graph database (neo4j) and other database systems [15][10].

The results also reveal that the graph queries formulated are very close to natural languages stated in the multi-perspective inquiries. For university managers or leaders who are often in the middle of meetings where they are requested to provide random and ad-hoc information regarding their universities, graph queries' capabilities will undoubtedly be beneficial.

In addition to reporting and analytical features supported by relational databases and queries, which most Executive Information Systems (EIS) currently have, organizations may need to develop graphical and analytical capabilities backed up by graph databases and queries.

Previous studies by [16], [17] added chatbot capabilities to existing EIS so that executives can view and get information about their universities by commanding through voices and texts. The problems with those systems are they do not know beyond the specific knowledge embedded in them. The commands are limited because they and their associated SQL queries must be pre-defined first. So, those systems have not addressed ad hoc, random requests for information.

The graph database and queries have great possibilities to overcome that problem. They can understand and produce intuitive results despite blind questions, which seem hard to comprehend and look meaningless, as shown in case 1 and case2. The *node-relationship-node* will arguably become the key in interpreting random requests for information as long as we can map things to nodes in the database. Indeed, the implementation will remain a challenge, especially when integrating data from different databases, often legacy systems, where documentations are less available or do not exist.

IV. Conclusion

The research graph database presented in this study has shown to efficiently answer the inquiries compared to the relational database and visually provide engaging discernment.

However, the research reported in this paper is still in the early stage and limited. Therefore, in the future works, the number of use cases covered, variety of data to be integrated, and analytical capabilities need to be increased and further explored to see graph database full potentials for universities' benefit.

Acknowledgment

Authors would like to thank Lembaga Penelitian dan Pengabdian Masyarakat – Universitas Pembangunan Nasional Veteran Jawa Timur for funding this research.

References

[1] H. Lu, Z. Hong, and M. Shi, “Analysis of film data based on Neo4j,” in Proceedings - 16th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2017, 2017.

[2] Y. Ma, Z. Wu, L. Guan, B. Zhou, and R. Li, “Study on the relationship between transmission line failure rate and lightning information based on Neo4j,” in POWERCON 2014 - 2014 International Conference on Power System Technology: Towards Green, Efficient and Smart Power System, Proceedings, 2014.

[3] S. Shaikh, S. Rathi, and P. Janrao, “Recommendation system in e-commerce websites: a graph based approached,” in Proceedings - 7th IEEE International Advanced Computing Conference, IACC 2017, 2017.

[4] V. Bajaj, R. B. Panda, C. Dabas, and P. Kaur, “Graph database for recipe recommendations,” in 2018 7th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions, ICRITO 2018, 2018.

[5] I. N. P. W. Dharmawan and R. Sarno, “Book recommendation using Neo4j graph database in BibTeX book metadata,” in Proceeding - 2017 3rd International Conference on Science in Information Technology: Theory and Application of IT for Education, Industry and Society in Big Data Era, ICSITech 2017, 2017.

[6] N. Yi, C. Li, X. Feng, and M. Shi, “Design and implementation of movie recommender system based on graph database,” in Proceedings - 2017 14th Web Information Systems and Applications Conference, WISA 2017, 2018.

[7] T. Konno, R. Huang, T. Ban, and C. Huang, “Goods recommendation based on retail knowledge in a Neo4j graph database combined with an inference mechanism implemented in jess,” in 2017 IEEE SmartWorld Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation, 2018.

[8] Y. Zhu, E. Yan, and I. Y. Song, “The use of a graph-based system to improve bibliographic information retrieval: System design, implementation, and evaluation,” J. Assoc. Inf. Sci. Technol., 2017.

[9] D. Hristovski, A. Kastrin, D. Dinevski, and T. C. Rindflesch, “Constructing a graph database for semantic literature-based discovery,” in Studies in Health Technology and Informatics, 2015.

[10] B.-H. Yoon, S.-K. Kim, and S.-Y. Kim, “Use of graph database for the integration of heterogeneous biological data,” Genomics Inform., vol. 15, no. 1, p. 19, 2017.

[11] B. M. Sasaki, “The 5-minute interview: David Meza, Chief Knowledge Architect, NASA,” 2016.

[12] M. Boeker, W. Vach, and E. Motschall, “Google Scholar as replacement for systematic literature searches: good relative recall and precision are not enough,” BMC Med. Res. Methodol., 2013.

[13] E. D. López-Cózar, N. Robinson-García, and D. Torres-Salinas, “The google scholar experiment: How to index false papers and manipulate bibliometric indicators,” J. Assoc. Inf. Sci. Technol., 2014.

[14] A. S. Ahmar et al., “Lecturers’ understanding on indexing databases of SINTA, DOAJ, Google Scholar, SCOPUS, and Web of Science: A study of Indonesians,” J. Phys. Conf. Ser., vol. 954, pp. 0–17, 2018.

[15] M. Sharma, V. D. Sharma, and M. M. Bundele, “Performance analysis of RDBMS and No SQL databases: PostgreSQL, MongoDB and Neo4j,” in 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering, ICRAIE 2018, 2018.

[16] M. I. Afandi, E. D. Wahyuni, and S. Mukaromah, “Mobile business intelligence assistant (m-BELA) for higher education executives,” 2019 4th Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2019.

[17] M. I. Afandi and E. D. Wahyuni, “Prototype of voice commanded university executive business intelligence assistant (BELA),” vol. 1, no. ICST, pp. 4–8, 2018.