DataNexus: Uma Ferramenta para Auxiliar a Análise da Modelagem de Bancos de Dados Extensos*

Gustavo Moraes[1], **Victor Misael B.F. Carneiro**[1], **Angelo Brayner**[1]

[1]Departamento de Computação – Universidade Federal do Ceará (UFC)
Campus do Pici – Fortaleza – CE – Brasil

{gustavo.moraes, victor.misael}@alu.ufc.br, brayner@dc.ufc.br

Abstract. *Databases are becoming increasingly complex, with large volumes of data and tables. Enterprise environments often deal with hundreds or even thousands of tables, which makes visualizing and understanding their schemas challenging. This paper presents a tool that uses network science techniques, such as community detection and degree centrality, to provide a clear and manageable view of complex database schemas. The main objective is to facilitate the understanding of these schemas, contributing to more efficient management of database catalogs and supporting reverse engineering processes when poorly designed schemas negatively impact performance.*

Resumo. *Bancos de dados estão se tornando cada vez mais complexos, com grandes volumes de dados e tabelas. Ambientes corporativos frequentemente lidam com centenas ou até milhares de tabelas, o que torna desafiadora a visualização e a compreensão de seus esquemas. Este artigo apresenta uma ferramenta que utiliza técnicas de ciência das redes, como detecção de comunidades e centralidade de grau, para oferecer uma visão clara e gerenciável de esquemas de bancos de dados complexos. O objetivo principal é facilitar a compreensão desses esquemas, contribuindo para a gestão mais eficiente de catálogos de bancos de dados e suporte a processos de engenharia reversa, quando os esquemas mal projetados impactam negativamente na performance.*

1. Introdução

Bancos de dados têm se tornado cada vez mais complexos, com grandes volumes de dados e com esquemas difíceis de serem compreendidos, devido ao elevado número de tabelas e alto grau de relacionamentos entre estas. Adicionalmente, bancos de dados são geralmente mal documentados. Tal cenário, inviabiliza a ”leitura”(visualização) e análise de esquemas de bancos de dados complexos, por parte de administradores de bancos de dados (DBA, acrônimo do inglês) experientes.

É fato que a gestão de bancos de dados complexos é crucial para o funcionamento eficiente das organizações em diversos setores. Contudo, as ferramentas existentes, para representar graficamente esquemas de bancos de dados, dificultam uma visualização de fácil entendimento no caso de bancos de dados com centenas ou até milhares de tabelas. Essas ferramentas frequentemente geram diagramas com uma rede densa de nós e conexões. Sem soluções adequadas, essa tarefa pode ser demorada e propensa a interpretações equivocadas ou incompletas do banco de dados.

*Video da demonstração da ferramenta disponível em: https://www.youtube.com/watch?v=euX4X1IsGUA

A aplicação de algoritmos de detecção de comunidades, oriundos da área de ciência das redes [Lancichinetti and Fortunato 2009], representa uma abordagem inovadora para enfrentar os desafios mencionados anteriormente. Esses algoritmos agrupam tabelas que têm interações significativas entre si. Por exemplo, em um esquema corporativo, pode-se encontrar áreas da aplicação intercaladas, onde tabelas relacionadas a setores específicos, como recursos humanos ou financeiro, tendem a ter mais conexões (relacionamentos) entre si. A detecção desses agrupamentos proporciona uma visão mais abstrata e gerenciável de grandes bancos de dados. Esses agrupamentos podem representar diferentes sub-esquemas dentro de um banco de dados complexo. Para ilustrar essa ideia de sub-esquema (ou contexto semântico), pode-se imaginar um banco de dados de uma empresa que armazena dados de todas ás áreas da empresa com um único esquema. Com o conceito de agrupamentos, pode-se identificar os diversos sub-esquemas, como recursos humanos e financeiro, correspondentes aos dados de diferentes setores da empresa.

Este trabalho apresenta, portanto, uma ferramenta destinada a auxiliar na compreensão de esquemas de bancos de dados complexos, utilizando técnicas de detecção de comunidades para agrupar tabelas. A ferramenta também permite a análise de aspectos específicos do banco de dados, como a identificação de tabelas com maior influência (maior número de chaves estrangeiras), maior prestígio (mais referenciadas por outras tabelas) e maior número de tuplas ou índices. Estas análises fornecem *insights* valiosos que podem melhorar a eficiência na gestão e utilização dos dados.

Este artigo está estruturado da seguinte forma: na Seção 2, discutimos os trabalhos relacionados. A Seção 3, apresenta a ferramenta DataNexus, incluindo a sua arquitetura e funcionamento. Finalmente, na Seção 4, concluímos o trabalho e sugerimos direções futuras para continuar o desenvolvimento da ferramenta.

2. Trabalhos Relacionados

Ferramentas populares de banco de dados como DBeaver, pgAdmin, DataGrip e MySQL Workbench são amplamente usadas por DBAs devido às suas funcionalidades e interfaces intuitivas. No entanto, essas ferramentas enfrentam limitações ao visualizar grandes e complexos esquemas de banco de dados, gerando grafos densos que dificultam a compreensão, especialmente para quem está interagindo com a modelagem pela primeira vez.

A detecção de comunidades em ciência das redes pode ser definida como a identificação de subgrupos coesos de nós mais conectados entre si do que com o restante da rede, facilitando a compreensão de sistemas complexos. Diversos algoritmos específicos são propostos para diferentes tipos de redes [Lancichinetti and Fortunato 2009]. Por exemplo, no trabalho de [Bedi and Sharma 2016], investiga-se a aplicação da detecção de comunidades em redes sociais. Esse tipo de rede pode ser representado por um grafo, onde os nós representam pessoas e as arestas representam interações entre elas (seguidores, amigos, etc.). A ferramenta H-BOLD [Desimoni et al. 2020] usa a detecção de comunidades para facilitar a visualização e exploração de grandes conjuntos de dados vinculados (*Linked Data*) publicados via endpoints SPARQ. Em sistemas complexos como esses, a detecção de comunidades pode ser benéfica para inúmeras aplicações, como a identificação de grupos com interesses semelhantes.

Em [Ledesma González et al. 2021], foi realizado um estudo em três cidades da Espanha, usando uma rede de destinos turísticos. Cada nó representa um destino turístico,

e as arestas representam relações entre esses destinos, como dependências e benefícios mútuos. O estudo focou em usar a centralidade de grau para entender como os nós se relacionam na rede. Foi possível medir o prestígio (*indegree*) e a influência (*outdegree*) dos nós. O prestígio de um nó refere-se ao número de nós que apontam para ele. Já a influência de um nó é o quanto ele influencia o prestígio dos outros nós na rede.

Portanto, técnicas detecção de comunidades, análise de prestígio e influência, podem ser aplicadas para aprimorar a compreensão de esquemas de bancos de dados complexos. Por isso, a DataNexus oferece uma visão mais clara e organizada de esquemas de bancos, facilitando a análise e gestão de esquemas de banco de dados extensos, proporcionando uma compreensão mais profunda das inter-relações entre os dados.

3. DataNexus: Uma Estratégia Inovadora para Visualização de esquemas de Bancos de Dados

Nesta seção, serão apresentadas as funcionalidade e propriedades da ferramenta proposta.

3.1. Arquitetura da DataNexus

A Figura 1 apresenta a arquitetura da DataNexus que segue um modelo cliente-servidor, com o serviço back-end implementado em Python[1]. Este back-end é integrado a diversas bibliotecas, incluindo drivers de conexão com os bancos de dados (atualmente suportando PostgreSQL e MySQL) e técnicas de ciência de redes. O usuário interage com uma aplicação web front-end que renderiza os grafos usando a biblioteca JavaScript D3. Além disso, a ferramenta oferece recursos de cache para os modelos de dados analisados, armazenando-os tanto em memória quanto em arquivos temporários. Essa abordagem permite reduzir significativamente a necessidade de consultas constantes aos bancos de dados alvo, melhorando a performance da aplicação.

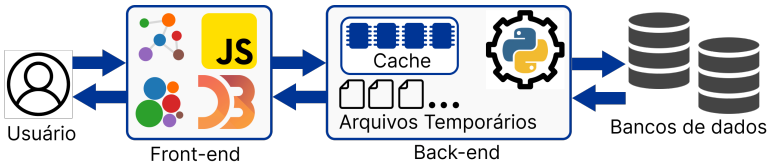

Figura 1. Arquitetura da DataNexus

3.2. A Interface da DataNexus

A Figura 2 apresenta a interface gráfica da DataNexus, composta por três painéis verticais. No painel à esquerda **(1)**, encontramos o menu principal, onde o usuário pode criar novas conexões clicando no botão *New Connection* **(1.1)**. Um modal será aberto, permitindo adicionar uma nova conexão inserindo as credenciais de acesso. Em **(1.2)**, o menu exibe as opções de visualização. O usuário pode definir o tamanho dos nós (tabelas) com base no número de relacionamentos, tuplas ou índices da tabela. Adicionalmente, ao escolher a opção de relacionamentos, é possível selecionar entre três critérios específicos de relacionamento: prestígio, influência ou ambos (ver Seção 3.3). Finalmente, o usuário pode optar por diferentes layouts de visualização: Comunidades concentradas (*Concentrated Community Graph*) ou comunidades espalhadas (*Spread Community Graph*) (ver Seção 3.4). Estas opções proporcionam flexibilidade na forma como as tabelas serão apresentadas, permitindo uma análise mais direcionada conforme a necessidade do usuário.

[1]Código da ferramenta disponível em: https://github.com/ggustavo/DataNexus

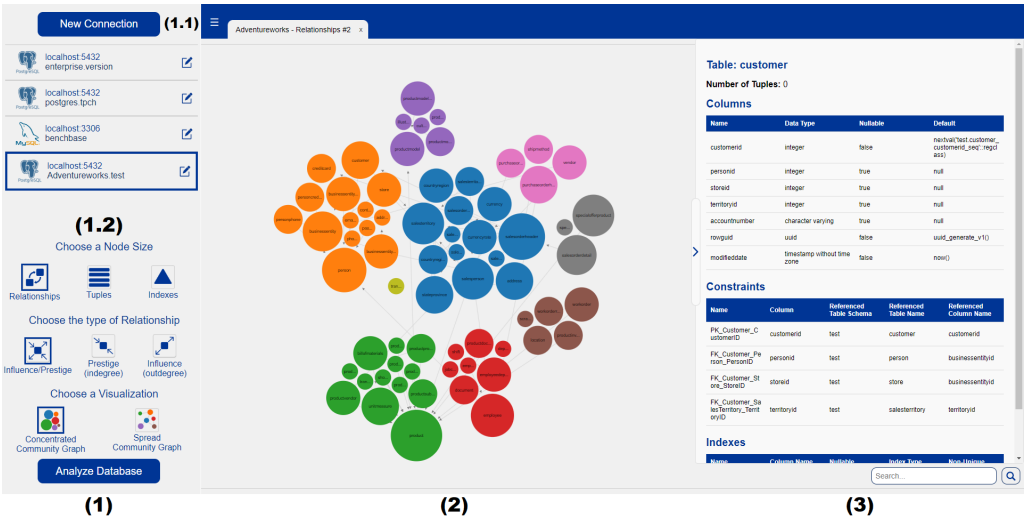

Figura 2. Interface gráfica da ferramenta.

O painel central **(2)** é dedicado à visualização do modelo de dados, onde as cores representam as comunidades de tabelas e as arestas mostram os relacionamentos entre elas. A ferramenta renderiza grafos de forma interativa, exibindo a estrutura e as conexões entre as tabelas, permitindo ao usuário mover e dar zoom na visualização. O painel à direita **(3)** exibe os metadados das tabelas selecionadas. Ao clicar em uma tabela no grafo, seus detalhes são apresentados, incluindo o número de tuplas, colunas, tipos de dados, restrições e índices. Para otimizar a área de visualização e facilitar a análise, os painéis laterais **(1)** e **(3)** podem ser colapsados, permitindo que o grafo ocupe toda a tela.

3.3. Tipos de Visualização dos Nós

A DataNexus possibilita a visualização de diferentes propriedades do banco de dados, como quantidade de índices e cardinalidade de cada tabela.

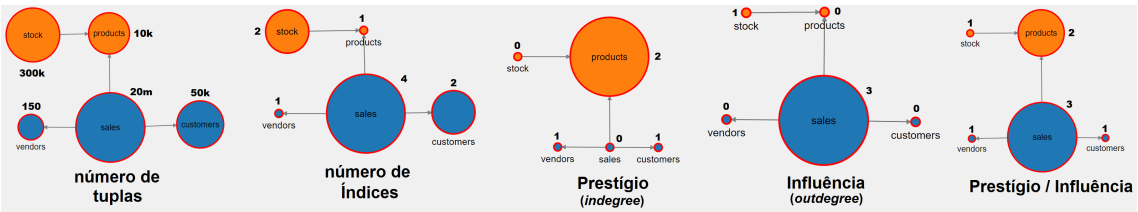

Figura 3. Diferentes formas de visualização de nós.

A Figura 3 apresenta exemplos de diferentes formas de visualização dos tamanhos dos nós na ferramenta DataNexus. Cada visualização oferece uma perspectiva distinta, facilitando a compreensão das tabelas dentro do modelo de dados. Abaixo, detalhamos os benefícios de visualizar o número de tuplas, índices, prestígio e influência de cada tabela.

Número de tuplas. Visualizar o número de tuplas de cada tabela ajuda a identificar rapidamente quais tabelas contêm mais dados. Isso é crucial para a otimização de consultas e gestão de performance, pois tabelas com um grande volume de tuplas podem se tornar gargalos em operações de leitura e escrita.

Número de índices. Exibir o número de índices de cada tabela permite identificar quais tabelas têm mais mecanismos de busca e ordenação. Índices são essenciais para melhorar o tempo das consultas, mas seu uso excessivo pode impactar negativamente a performance de operações de inserção e atualização. Dessa forma, visualizar o número de índices auxilia na análise de *trade-offs* entre velocidade de leitura e desempenho de escrita.

Prestígio (*indegree*). Tabelas com alto prestígio são aquelas que são referenciadas por muitas outras tabelas. Identificar essas tabelas é vantajoso porque atualizações ou remoções nessas tabelas podem causar operações em cascata em todas as tabelas que as referenciam. Esse aspecto é crucial no contexto transacional, pois alterações em tabelas de alto prestígio podem afetar o desempenho de várias transações simultâneas.

Influência (*outdegree*). Tabelas influentes são aquelas que referenciam diversas outras tabelas, representando pontos-chave de interconexão no modelo de dados. Identificar e melhorar a eficiência dessas tabelas é fundamental para acelerar operações complexas de junção em consultas que envolvem múltiplas tabelas, o que beneficia diretamente o desempenho do sistema em operações transacionais de alta complexidade.

3.4. *Layouts* de Visualização do Grafo e Detecção de Comunidades

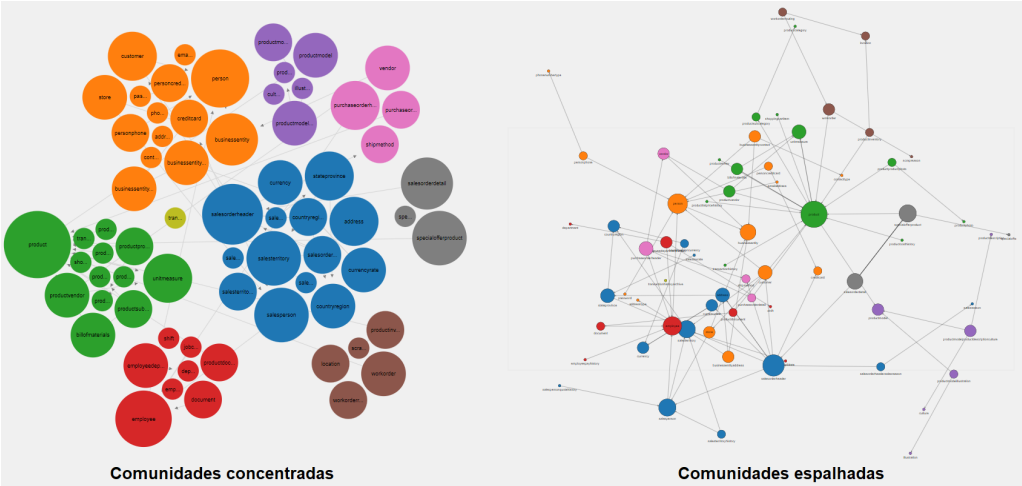

Figura 4. *Layouts* de visualização do grafo

A Figura 4 apresenta o mesmo esquema de banco de dados visualizado de duas maneiras diferentes. As cores em ambos os layouts foram atribuídas pelo algoritmo de detecção de comunidades *greedy modularity maximization* [NetworkX 2023], que começa considerando cada nó como uma comunidade. Em seguida, o algoritmo combina iterativamente as comunidades para maximizar a modularidade da rede, uma medida que avalia a qualidade da divisão dos nós. O layout de comunidades concentradas visa posicionar as tabelas que pertencem à mesma comunidade próximas umas das outras. Por outro lado, o layout de comunidades espalhadas adota uma abordagem mais livre na disposição das tabelas, similar às ferramentas tradicionais, embora sem os atributos visuais detalhados, os quais podem ser consultados no painel à direita **(3)** (ver Seção 3).

Na Figura 5, comparamos a visualização de um esquema corporativo real, composto por 135 tabelas, gerado pelo pgAdmin (à esquerda) e pela DataNexus[2] (à direita). A

[2]Devido à política de privacidade, os nomes das tabelas foram anonimizados.

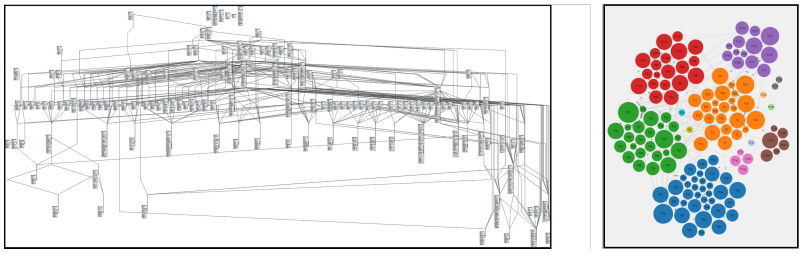

Figura 5. Comparação de diagramas: pgAdmin vs DataNexus

visualização pelo pgAdmin segue um formato tradicional, mostrando uma rede densa de tabelas e suas conexões. Embora essa abordagem forneça uma visão detalhada, incluindo os atributos das tabelas, a complexidade visual torna difícil interpretar as relações. Em contraste, a visualização gerada pela DataNexus organiza as tabelas usando algoritmos de detecção de comunidades. Essa técnica agrupa as tabelas em comunidades, o que facilita a navegação e a análise do modelo de dados. Em vez de analisar um grafo denso e complexo, o usuário pode se concentrar em uma comunidade de cada vez, entendendo melhor as interações internas antes de explorar outras partes do modelo.

4. Conclusão

Este trabalho introduziu a DataNexus, uma ferramenta projetada para auxiliar na análise de modelos extensos de bancos de dados que utiliza técnicas de ciência das redes. Exploramos variações nas visualizações de tamanhos de nós e layouts para facilitar a compreensão dos modelos de dados. Além disso, utilizamos algoritmos de detecção de comunidades para agrupar tabelas relacionadas, proporcionando uma visão macro aprimorada. Para futuras melhorias, planejamos incorporar técnicas adicionais de análise de redes. Também pretendemos integrar à ferramenta métodos para identificar automaticamente anomalias nos modelos de dados, proporcionando *insights* para otimizações. Outro objetivo é a conversão da ferramenta para um *plugin* capaz de integrar com ferramentas populares como o DBeaver. Esta integração permitirá aos usuários aproveitar as capacidades da DataNexus diretamente em suas plataformas de trabalho habituais.

Referências

Bedi, P. and Sharma, C. (2016). Community detection in social networks. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 6(3):115–135.

Desimoni, F., Po, L., et al. (2020). Providing effective visualizations over big linked data. In *EDBT/ICDT 2020 Workshops*. CEUR-WS.

Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 80(5):056117.

Ledesma González, O., Merinero-Rodríguez, R., and Pulido-Fernández, J. I. (2021). Tourist destination development and social network analysis: What does degree centrality contribute? *International Journal of Tourism Research*, 23(4):652–666.

NetworkX (2023). Communities. `https://networkx.org/documentation/stable/reference/algorithms/community.html`. [Accessed on 1th September 2024].