

PFKG-LLM: Exploração Guiada de Dados Fiscais em Larga Escala com Grafos de Conhecimento e LLMs

Roberval G. Mariano¹, Vania P. Vidal¹

¹Departamento de Ciência da Computação – Universidade Federal do Ceará (UFC)–
60.020-181 – Fortaleza – CE – Brazil

mariano.rgm.2015@gmail.com, vaniap.vidal@gmail.com

Abstract. *Fiscal data analysis in Big Data environments requires identifying which entities should be prioritized among large and heterogeneous datasets. Traditional auditing approaches become insufficient in this context. This work presents **PFKG-LLM**, a platform based on Knowledge Graphs and LLM agents for guided exploration of fiscal data. The system models data as a large-scale graph and uses metrics to iteratively select relevant entities. Based on these analyses, new candidates are dynamically identified, enabling an adaptive auditing process. The platform integrates Graph RAG, interactive visualizations and natural language queries translated into SPARQL. Results indicate a scalable and knowledge-driven approach for fiscal data analysis.*

Resumo. *A análise de dados fiscais em Big Data exige identificar entidades prioritárias em grandes volumes de dados heterogêneos. Abordagens tradicionais tornam-se insuficientes nesse cenário. Este trabalho apresenta o **PFKG-LLM**, uma plataforma baseada em Grafos de Conhecimento e agentes LLM para exploração guiada por dados. A solução modela os dados como um grafo em larga escala e utiliza métricas para selecionar iterativamente entidades relevantes. A partir dessas análises, novos candidatos são identificados dinamicamente, promovendo um processo adaptativo. A plataforma integra Graph RAG, visualizações interativas e consultas em linguagem natural convertidas em SPARQL. Os resultados indicam uma abordagem direcionada, escalável e orientada ao conhecimento.*

1. Introdução

A análise de dados fiscais em ambientes de Big Data representa um desafio crescente para auditoria e fiscalização tributária. O ecossistema do Sistema Público de Escrituração Digital (SPED) integra bilhões de documentos fiscais eletrônicos e informações heterogêneas provenientes de múltiplas fontes, tornando inviável a exploração manual ou baseada apenas em consultas estáticas. Nesse cenário, um dos principais desafios consiste em identificar, de forma dinâmica, quais entidades e relações devem ser priorizadas durante o processo analítico.

Grafos de Conhecimento (Knowledge Graphs – KG) têm sido utilizados para representar dados fiscais e suas relações semânticas, permitindo consultas estruturadas e análise relacional em larga escala. Paralelamente, Large Language Models (LLMs) ampliaram as possibilidades de interação com dados complexos, especialmente em tarefas envolvendo consultas em linguagem natural e exploração analítica assistida.

Contudo, ainda existe uma lacuna na integração entre Graph RAG, consultas semânticas e mecanismos iterativos de seleção de entidades voltados à análise fiscal.

Este trabalho apresenta o PFKG-LLM, uma plataforma de exploração guiada de dados fiscais baseada em Enterprise Knowledge Graphs (EKG), SPARQL e agentes LLM. A plataforma combina métricas analíticas, consultas semânticas e geração automática de consultas para apoiar a identificação iterativa de entidades relevantes em grafos fiscais de larga escala. O processo utiliza agentes LLM para interpretar o contexto analítico, gerar consultas SPARQL e expandir dinamicamente a exploração do grafo a partir dos resultados obtidos.

A aplicação integra exploração visual, consultas SPARQL e Graph RAG para análise iterativa de candidatos fiscais. A abordagem permite reduzir a dependência de heurísticas fixas e direcionar a análise para regiões mais relevantes do grafo, favorecendo a descoberta progressiva de padrões e relações não triviais.

Como contribuições, destacam-se: (i) uma plataforma de Graph RAG fiscal baseada em Grafos de Conhecimento e agentes LLM; (ii) um processo iterativo de exploração guiada para seleção dinâmica de entidades relevantes; (iii) a integração entre consultas em linguagem natural, SPARQL e visualização interativa de grafos fiscais em larga escala.

2. Trabalhos Relacionados

Grafos de Conhecimento (Knowledge Graphs – KG) têm sido amplamente utilizados na integração e exploração de dados heterogêneos, permitindo representação semântica explícita de entidades e relações, bem como consultas estruturadas sobre dados altamente conectados, HOGAN et al. (2021). Em ambientes organizacionais, arquiteturas baseadas em Enterprise Knowledge Graphs (EKG) permitem consolidar múltiplas fontes e perspectivas analíticas em uma estrutura semântica unificada, VIDAL et al. (2026).

Abordagens baseadas em grafos têm sido empregadas na detecção de anomalias e fraudes em cenários financeiros e fiscais, AKOGLU, TONG, KOUTRA (2014), WEBER et al. (2019) e POURHABIB et al. (2020). Essas abordagens exploram propriedades estruturais e relacionais, porém normalmente dependem de consultas previamente definidas ou pipelines analíticos rígidos.

Mais recentemente, Large Language Models (LLMs) passaram a ser integrados a Grafos de Conhecimento em abordagens como Graph RAG, permitindo consultas em linguagem natural, geração automática de consultas formais e exploração assistida de dados estruturados, LEWIS et al. (2020), PAN et al. (2023), PENG et al. (2024) e GAO et al. (2023).

Entretanto, a maioria dessas abordagens permanece centrada em recuperação de informação ou suporte à consulta, sem tratar explicitamente a seleção iterativa de entidades relevantes e a expansão adaptativa da análise em grafos fiscais de larga escala. O PFKG-LLM posiciona-se nessa lacuna ao integrar Enterprise Knowledge Graphs, consultas SPARQL, métricas analíticas e agentes LLM em uma plataforma de exploração guiada e iterativa voltada à análise fiscal.

3. Arquitetura e Processo de Exploração Guiada da Plataforma

3.1 Arquitetura da Plataforma

O PFKG-LLM foi desenvolvido como uma plataforma de exploração guiada de dados fiscais baseada na integração entre Enterprise Knowledge Graphs (EKG), consultas SPARQL e agentes LLM. A aplicação utiliza um pipeline semântico para representação e exploração de grandes volumes de dados fiscais estruturados como grafos de conhecimento.

A arquitetura da plataforma é composta por:

- (i) camada de armazenamento semântico baseada em RDF e SPARQL;
- (ii) repositório de grafos em GraphDB;
- (iii) agentes LLM para geração e refinamento de consultas;
- (iv) módulo de exploração visual de grafos; e
- (v) interface interativa para análise e navegação dos dados.

Os dados fiscais são organizados segundo a arquitetura EKG proposta por Vidal et al. (2026), incluindo camadas de Exported Views, Linkset Views, Unification Views e Fusion Views, permitindo integração semântica e consolidação progressiva das entidades e relações.

A plataforma foi implementada para execução local utilizando Streamlit, GraphDB, RDF/SPARQL, CrewAI, Ollama e bibliotecas de visualização interativa de grafos.

3.2 Processo de Exploração Guiada

O processo de exploração inicia com a seleção de entidades relevantes a partir de métricas estruturais e agregadas calculadas sobre o grafo fiscal. Essas métricas incluem conectividade, frequência de transações e valores fiscais agregados.

As entidades priorizadas são utilizadas como ponto de partida para consultas iterativas sobre o grafo. A partir dos resultados obtidos, novas entidades e relações relevantes são incorporadas ao processo analítico, permitindo expansão progressiva da exploração.

3.3 Agentes LLM e Graph RAG

Os agentes LLM atuam na interpretação do contexto analítico e na geração de consultas sobre o grafo de conhecimento. A plataforma utiliza uma abordagem Graph RAG, na qual consultas em linguagem natural são convertidas em consultas SPARQL executadas sobre o repositório semântico.

Os agentes são responsáveis por: (i) interpretar perguntas em linguagem natural; (ii) gerar e refinar consultas SPARQL; (iii) analisar os resultados retornados pelo grafo; e (iv) apoiar a identificação iterativa de novos candidatos relevantes.

Essa abordagem permite reduzir a dependência de consultas estáticas e favorecer exploração adaptativa baseada em evidências extraídas diretamente do grafo.

3.4 Pipeline de Execução

A Figura 1 apresenta o pipeline completo da plataforma. O processo inicia com a carga e organização semântica dos dados fiscais no EKG. Em seguida, métricas analíticas são calculadas para seleção de entidades prioritárias.

Os agentes LLM geram consultas SPARQL a partir do contexto analítico e dos resultados obtidos. As respostas retornadas pelo grafo alimentam novas iterações de exploração, permitindo expansão progressiva da análise e descoberta de relações relevantes em larga escala.

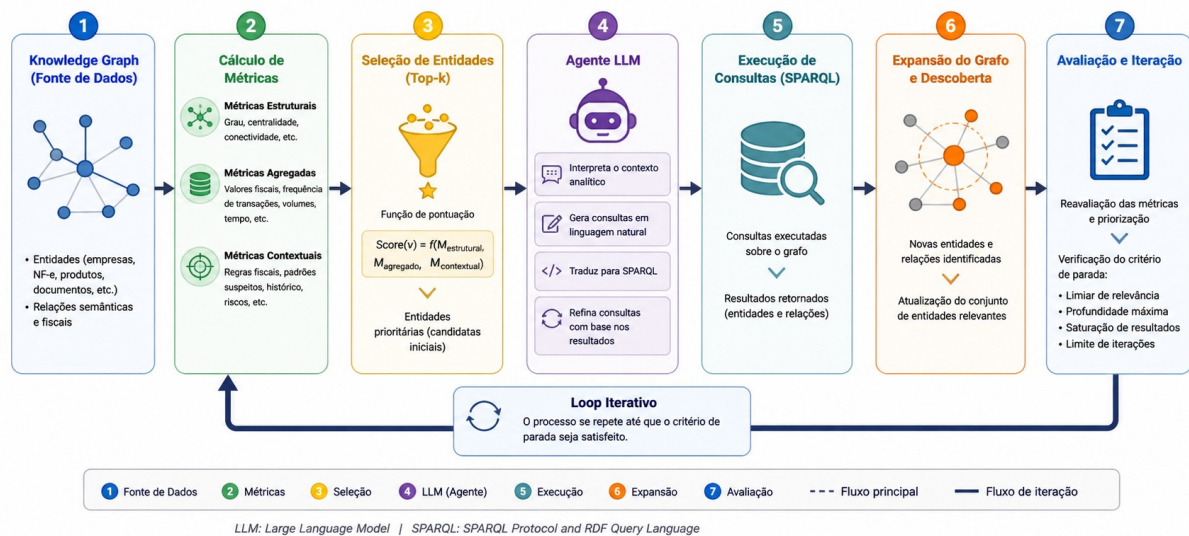


Figura 1. Pipeline de Exploração Guiada da Plataforma PFKG-LLM.

4. Demonstração da Aplicação e Cenário de Uso

4.1 Cenário de Uso

A plataforma foi aplicada em um cenário de exploração de dados fiscais estruturados como Grafos de Conhecimento, considerando documentos fiscais eletrônicos, entidades empresariais e relações transacionais representadas semanticamente em RDF. O processo de análise inicia com a seleção de entidades prioritárias a partir de métricas estruturais e agregadas calculadas sobre o grafo.

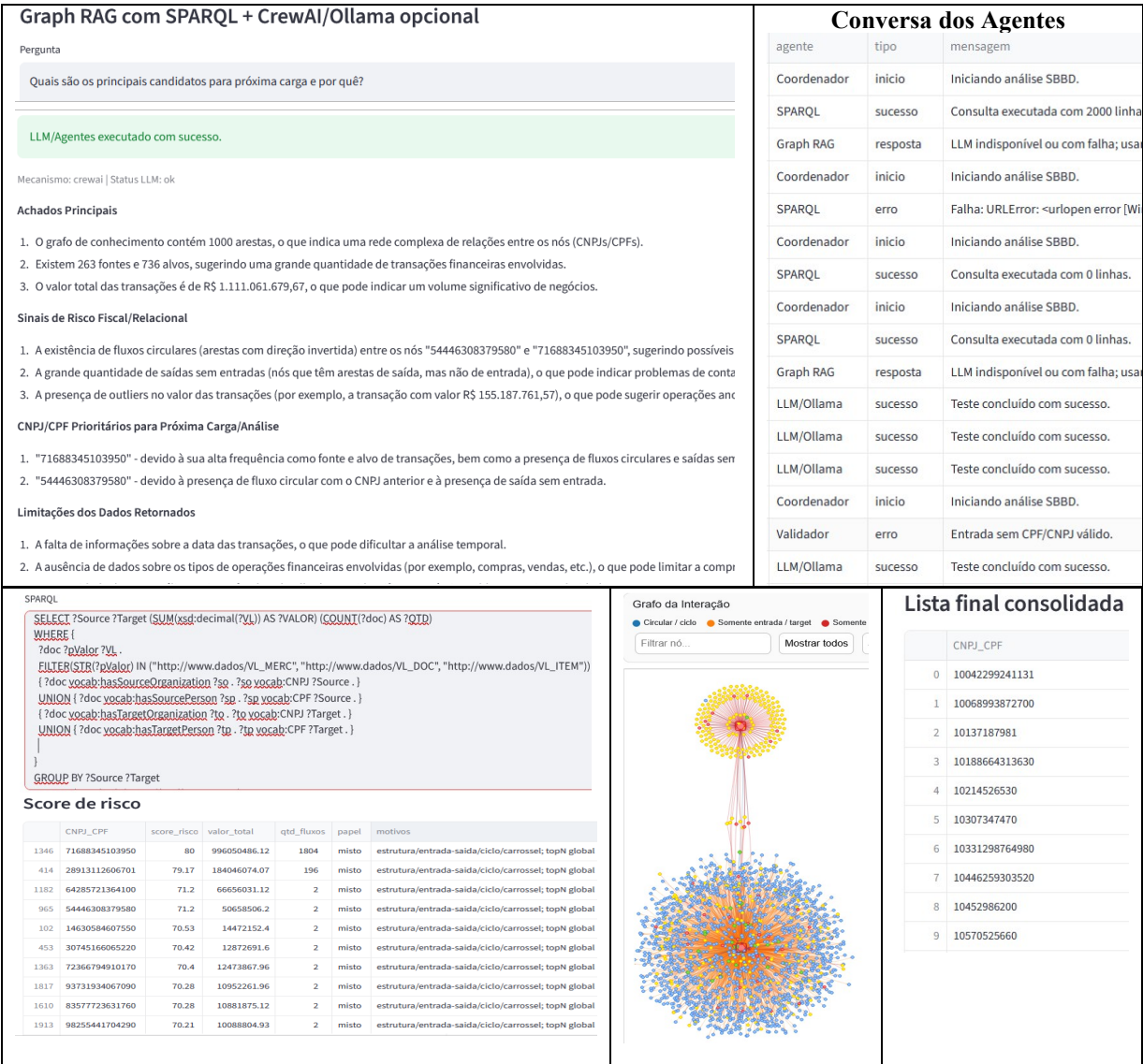
A partir dessas entidades, o usuário pode realizar consultas em linguagem natural, explorar relações no grafo e expandir iterativamente a análise para identificação de novos candidatos relevantes. O processo permite direcionar a exploração para regiões mais relevantes do grafo, reduzindo a dependência de consultas estáticas e análise manual.

4.2 Exploração Interativa e Graph RAG

A aplicação integra uma interface interativa para exploração visual do grafo fiscal e consultas assistidas por agentes LLM. O usuário pode realizar perguntas em linguagem natural, as quais são convertidas automaticamente em consultas SPARQL executadas sobre o repositório semântico.

Os resultados retornados pelo grafo são utilizados pelos agentes para refinamento iterativo das consultas e expansão progressiva da análise. A abordagem permite combinar recuperação semântica baseada em grafos (Graph RAG), exploração visual e geração automática de consultas em um mesmo ambiente analítico.

Durante a demonstração, o usuário pode realizar consultas em linguagem natural, acompanhar a geração de consultas SPARQL pelos agentes LLM e explorar dinamicamente as relações identificadas no grafo fiscal. A Figura 2 apresenta a interface da aplicação, incluindo exploração Graph RAG, consultas analíticas e navegação interativa no grafo fiscal, com: 2.158 arestas, 459 entidades Source, 1.692 entidades Target e R\$ 1.185.226.011,39 (10% dos dados reais ofuscados).



em linguagem natural; (iv) geração automática de consultas via agentes LLM; (v) seleção iterativa de candidatos relevantes; e (vi) exploração guiada baseada em métricas analíticas.

A aplicação foi desenvolvida para execução local, utilizando GraphDB, Streamlit, RDF/SPARQL, CrewAI e Ollama, permitindo exploração interativa de grandes volumes de dados fiscais estruturados semanticamente.

5. Conclusão e Trabalhos Futuros

Este trabalho apresentou o PFKG-LLM, uma plataforma para exploração guiada de dados fiscais baseada na integração entre Enterprise Knowledge Graphs, consultas SPARQL e agentes LLM. A abordagem combina Graph RAG, exploração visual e geração automática de consultas para apoiar a identificação iterativa de entidades relevantes em grafos fiscais de larga escala.

A aplicação permite integrar consultas em linguagem natural, exploração semântica e análise relacional em um ambiente unificado, reduzindo a dependência de consultas estáticas e favorecendo processos analíticos adaptativos orientados por evidências extraídas do grafo.

Como trabalhos futuros, destacam-se a ampliação das estratégias de priorização analítica, a investigação de mecanismos adicionais de explicabilidade para agentes LLM e a avaliação da abordagem em diferentes cenários de análise fiscal em larga escala.

Referências

- AKOGLU, Leman; TONG, Hanghang; KOUTRA, Danai. Graph-based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, v. 29, n. 3, p. 626–688, 2014. DOI: <https://doi.org/10.1007/s10618-014-0365-y>.
- GAO, Yunfan et al. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv*, 2023.
- HOGAN, Aidan et al. Knowledge graphs. *ACM Computing Surveys*, v. 54, n. 4, p. 1–37, 2021.
- LEWIS, Patrick et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *NeurIPS*, 2020.
- PAN, Shirui et al. Large language models and knowledge graphs: opportunities and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 2023.
- PENG, Boci et al. Graph Retrieval-Augmented Generation: A Survey. *arXiv:2408.08921v1 [cs.AI]* 15 Aug 2024
- POURHABIB, Amir et al. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, 2020.
- VIDAL, Vânia Maria Ponte et al. A Conceptual Framework for Building and Exploring Semantic Views of Enterprise Knowledge Graphs. *Journal of Information and Data Management (JIDM)*, 2026.
- WEBER, Mark et al. Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. *KDD*, 2019.