LATINOWARE 2025
22º Congresso Latino-americano de Software Livre e Tecnologias Abertas
Latin.Science 2025

Arquitetura Automatizada de Data Warehouse para Acesso ao Conhecimento Tecnológico em Bases de Patentes

Rodrigo Ramos Nogueira* , Alexandre Leopoldo Gonçalves*, Tobias Felipe Kiefer/
*Universidade Federal de Santa Catarina, Ibirama, Brasil.
rodrigo.nogueira@ifc.edu.br, a.l.goncalves@ufsc.br, tobias.f.kiefer@gmail.com

***Abstract*—This paper presents an automated Data Warehouse architecture designed for the analysis of USPTO patent data provided in XML format. The proposal includes a full ETL process with semantic enrichment and a multidimensional model, enabling OLAP queries and integration with Artificial Intelligence algorithms. Despite the limited dataset used, technical term analysis and neural network-based forecasting were performed, validating the model's potential to support technological knowledge management. Results indicate that the solution is applicable to competitive intelligence and the identification of emerging trends in innovation-driven environments.**

Keywords —Data Warehouse, ETL, Patents, OLAP, Artificial Intelligence, Knowledge Management, Dimensional Modeling, Technological Analysis.

***Resumo*— Este artigo apresenta uma arquitetura automatizada de Data Warehouse voltada à análise de patentes técnicas da USPTO, estruturadas em arquivos XML. A proposta inclui um processo completo de ETL com enriquecimento semântico e modelagem multidimensional, permitindo consultas OLAP e integração com algoritmos de Inteligência Artificial. Mesmo com uma amostra reduzida de dados, foram realizadas análises de termos técnicos e previsões com redes neurais, evidenciando o potencial do modelo em apoiar a gestão do conhecimento tecnológico. Os resultados demonstram a aplicabilidade da solução para inteligência competitiva e identificação de tendências emergentes em ambientes de inovação.**

***Palavras-chave* — Data Warehouse, ETL, Patentes, OLAP, Inteligência Artificial, Gestão do Conhecimento, Modelagem Dimensional, Análise Tecnológica.**

I. Introdução

Em um contexto de transformação digital acelerada, o acesso a informações técnicas atualizadas tornou-se essencial para a competitividade de empresas e instituições. As bases de patentes despontam como fontes organizadas e ricas em conhecimento técnico, permitindo o mapeamento do estado da arte em diversas áreas [1]. Além de seu papel legal, essas bases promovem a difusão tecnológica ao exigirem a descrição detalhada das inovações.

O valor estratégico das patentes vai além da proteção intelectual: elas impulsionam o desenvolvimento econômico, o empreendedorismo e a formulação de políticas públicas [2]. A análise dessas bases é, portanto, uma ferramenta relevante para a inteligência competitiva e para a gestão do conhecimento. No entanto, a extração de valor dessas informações exige ferramentas robustas, capazes de lidar com formatos complexos e volume massivo de dados [3]. A literatura aponta que documentos de patentes, especialmente os disponibilizados em XML pela USPTO, apresentam desafios típicos de Big Data, como recorrência temporal, heterogeneidade e grande volume [4]. Nesse cenário, arquiteturas baseadas em Data Warehouse surgem como solução adequada, ao integrar dados históricos, estruturá-los em modelos analíticos e viabilizar consultas multidimensionais [5].

Este trabalho propõe uma arquitetura automatizada para extração, transformação, carga (ETL) e análise de dados de patentes, organizada em três camadas: (i) um pipeline de ETL que processa arquivos XML brutos; (ii) um Data Warehouse com modelagem dimensional voltada à análise tecnológica; e (iii) uma interface de consumo analítico, com suporte a OLAP e integração com modelos de IA. A proposta visa apoiar decisões estratégicas e a geração de conhecimento em ambientes de inovação [6].

II. Trabalhos Correlatos

A análise de dados de patentes tem se consolidado como uma estratégia essencial para a geração de conhecimento tecnológico, promovendo avanços em inovação, competitividade institucional e políticas de desenvolvimento. As patentes, por sua natureza descritiva e técnica, oferecem uma das mais ricas fontes de informação científica acessível, sendo empregadas em estudos de prospecção tecnológica, mapeamento de redes de colaboração e detecção de tendências emergentes [2], [1].

Entretanto, o volume crescente e a complexidade estrutural dessas bases apresentam obstáculos relevantes para a análise automatizada. Um exemplo é o repositório da United States Patent and Trademark Office (USPTO), que semanalmente disponibiliza mais de 1GB de documentos no formato XML — um padrão semiestruturado que demanda técnicas avançadas de leitura, interpretação e extração [4]. Nesses cenários, o uso de pipelines de ETL (Extração, Transformação e Carga) torna-se indispensável, possibilitando o pré-processamento e organização dos dados

para análises posteriores.

Para apoiar esse fluxo de análise, arquiteturas baseadas em Data Warehouses (DWs) multidimensionais têm se mostrado altamente eficazes. A modelagem por esquemas estrela e o uso de consultas OLAP viabilizam a análise temporal, temática e geográfica de dados, organizando grandes volumes de informação de forma analítica e acessível [5].

A integração dessas estruturas com técnicas de Inteligência Artificial tem ampliado ainda mais o potencial dos estudos baseados em patentes. Modelos de aprendizado de máquina e redes neurais têm sido aplicados para sumarização automática de resumos [7], classificação temática [8], extração de palavras-chave [9], construção de grafos de conhecimento [10] e previsão de áreas tecnológicas emergentes [11]. Tais tarefas são viabilizadas por representações vetoriais, como embeddings, que permitem o processamento semântico dos textos técnicos por meio de arquiteturas generativas.

Estudos evidenciam o potencial do uso de pipelines em Python combinados a estruturas de Data Warehouse para o tratamento de dados complexos. González, Sakata e Nogueira [12] propõem uma solução de extração e carga de dados voltada para um Data Warehouse modelado em estrela no PostgreSQL, destacando a eficiência da estrutura multidimensional na organização de grandes volumes de dados e na execução de consultas analíticas — abordagem compatível com a proposta deste trabalho.

III. Desenvolvimento

O sistema proposto neste trabalho foi desenvolvido com o objetivo de automatizar todas as etapas envolvidas no processamento de dados de patentes, desde a extração dos arquivos XML da base da USPTO até a disponibilização dos dados estruturados para análise. A arquitetura foi organizada em três camadas principais: pipeline ETL, armazenamento em Data Warehouse e interface de acesso aos dados.

A Figura 1 apresenta o processo de ETL desenvolvido neste trabalho foi projetado para tratar os dados textuais e estruturados provenientes dos arquivos XML disponibilizados pela USPTO. A primeira etapa consiste na extração dos dados brutos, organizados em pacotes semanais com mais de 1GB, contendo descrições técnicas detalhadas de patentes. Esses arquivos apresentam alta complexidade hierárquica, o que exige parsing eficiente e tolerante a variações de estrutura. Após a extração, a fase de limpeza dos dados remove declarações duplicadas e inconsistências. Em seguida, os dados passam pela transformação, que inclui a conversão de datas (formato YYYYMMDD para dia/mês/ano), padronização de nomes, tokenização com expressões regulares e remoção de stopwords.

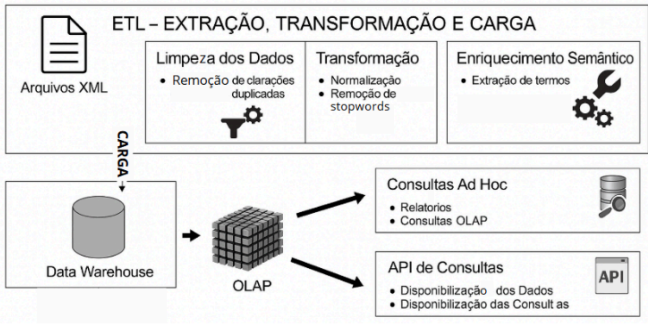

Fig2: Processo de ETL

Concluída a transformação, a etapa final corresponde ao enriquecimento semântico, em que técnicas de Processamento de Linguagem Natural e modelos LLM são aplicados para identificar termos técnicos relevantes nos resumos das patentes. Os dados resultantes são então carregados no Data Warehouse, onde ficam organizados para permitir consultas OLAP e geração de relatórios ad hoc. Como mostra a Figura 1, a arquitetura final ainda contempla uma API RESTful que fornece acesso estruturado às informações, permitindo sua integração com dashboards e modelos de IA. Essa estrutura modular garante flexibilidade, escalabilidade e a possibilidade de reaproveitamento em outros contextos de análise de dados técnicos.

A Figura 2 apresenta o modelo dimensional desenvolvido para o Data Warehouse de patentes, estruturado no formato estrela (star schema). No centro encontra-se a tabela fato fact_patents, que armazena as ocorrências de termos técnicos em patentes, com suas respectivas frequências (num_ocorr). Essa tabela é relacionada a seis dimensões: dim_words (termos extraídos dos resumos), dim_authors (nomes dos autores), dim_date (data de depósito), dim_countries (país de origem), dim_categories (subclasses tecnológicas) e dim_patents (metadados da patente). Essa modelagem permite análises multidimensionais ao longo do tempo, por país, autor ou setor tecnológico, viabilizando consultas OLAP otimizadas para fins de inteligência tecnológica e apoio à decisão estratégica.

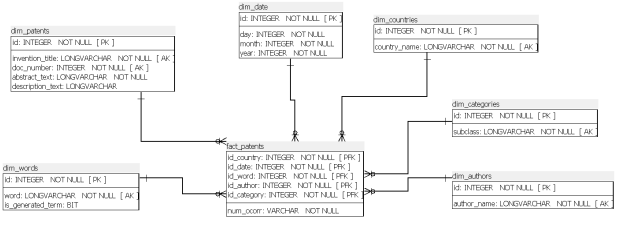

Fig2: Modelagem Multidimensional de Patentes

Uma vez os dados armazenados, foi desenvolvida uma interface de acesso via API RESTful, construída com FastAPI, permitindo a exposição dos dados de forma organizada e segura. A API possibilita filtros por autor, país, datas e termos específicos, servindo como base para a criação de dashboards e integração com modelos de Inteligência Artificial. Esse exemplo ilustra o consumo da API para consulta filtrada por autor e país, retornando os dados em formato JSON.

```
import requests
response = requests.get(
    "http://localhost:8000/patents",
    params={"author": "Einstein", "country": "US"}
)
print(response.json())
```

O conjunto de endpoints disponibilizados pela API RESTful, resumido no Quadro 1, oferece uma interface robusta e flexível para o acesso às informações processadas no Data Warehouse. Essas rotas permitem desde consultas simples — como a recuperação das palavras mais frequentes — até análises mais complexas de associação semântica e variação temporal. Essa estrutura é essencial para possibilitar a exploração automatizada e personalizada dos dados de patentes, facilitando a integração com ferramentas de visualização, sistemas de recomendação e modelos de Inteligência Artificial aplicados à prospecção tecnológica.

A API oferece um conjunto de endpoints para a análise de dados de patentes. Para a análise de palavras, o endpoint /words/top retorna as 30 palavras mais frequentes nos resumos. A frequência de palavras pode ser analisada de diferentes maneiras: /words/por-ano as agrupa por ano, /words/por-pais as organiza por país de origem e /words/por-autor as lista por autor. O endpoint /words/ranking-anual fornece o ranking das cinco palavras mais frequentes em cada ano. Para análises de coocorrência, a API disponibiliza o /words/associadas?termos=termo1,termo2, que retorna palavras associadas a termos específicos nos resumos, e o /words/associadas-tempo?termos=..., que realiza a mesma análise, mas com uma perspectiva temporal. Além disso, é possível buscar patentes vinculadas a um autor específico através do endpoint /authors/{nome}.

IV. Resultados Parciais

Os experimentos realizados com uma amostra de 15.399 patentes extraídas de dois arquivos XML da USPTO, datados de 29/05/2025 e 02/01/2020, revelaram que as seções da Classificação Internacional de Patentes (IPC) com maior volume de documentos são “Eletricidade” (4452 patentes), “Física” (4401) e “Necessidades Humanas” (2309). As principais subclasses incluem temas como Circuitos Eletrônicos, Medição e Testes, Instrumentação Científica e Cuidados Médicos, evidenciando a diversidade de técnicas e tecnologias nos dados analisados.

A Figura 3 apresenta um primeiro experimento com a visualização temporal da frequência de ocorrência dos principais termos técnicos identificados nas patentes. A estrutura multidimensional do *Data Warehouse* permitiu essa análise ao associar cada ocorrência de termo a diferentes dimensões, como o ano de depósito, o tipo de invenção e sua classificação. Por meio dessa abordagem, é possível observar padrões temporais e tendências emergentes, como o crescimento contínuo do termo *electronic device* a partir de 2022 e a oscilação acentuada de *pharmaceutical composition* ao longo dos anos. Essa capacidade de desagregar e cruzar informações em diferentes perspectivas é viabilizada pela modelagem dimensional adotada, reforçando a aplicabilidade da solução no suporte à gestão do conhecimento e à inteligência competitiva.

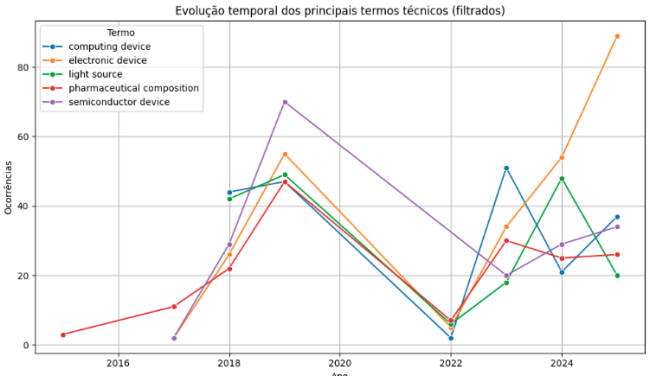

Fig2: Trend de termos em relação ao tempo

Em outro experimento realizado e reportado pela Figura 3 é mostrado uma evolução histórica e a previsão de ocorrências de termos técnicos em patentes com base em uma rede neural do tipo MLP (Multilayer Perceptron) que consome o modelo multidimensional. A abordagem adotada utiliza uma janela temporal de três anos para treinar modelos individuais por termo, aplicando uma previsão iterativa para os anos de 2026 a 2028. O modelo foi treinado com dados normalizados, aplicando a técnica de MinMaxScaler, o que garante maior estabilidade durante o treinamento. A etapa de previsão é feita de forma sequencial, permitindo que a saída de um ano seja usada como entrada para o seguinte, técnica comum em séries temporais.

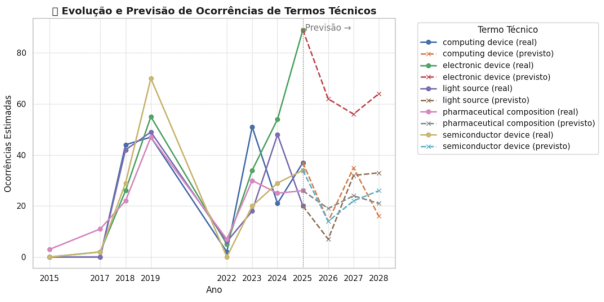

Fig3: Evolução e Previsão de Termos

Ao observar os resultados, nota-se uma projeção de crescimento relevante para o termo "electronic device", indicando sua consolidação como um eixo central da inovação tecnológica no período previsto. Termos como "pharmaceutical composition" e "light source" também apontam crescimento moderado, sugerindo continuidade em pesquisas na área de saúde e iluminação. Embora a base utilizada contenha um volume reduzido de dados, o objetivo principal foi validar a estrutura do modelo multidimensional proposto, demonstrando sua capacidade de alimentar algoritmos de IA. Ainda que os resultados devam ser interpretados com cautela, o experimento comprova a viabilidade de integrar análises preditivas ao Data Warehouse de patentes.

V. Conclusão

A arquitetura automatizada de Data Warehouse desenvolvida neste trabalho demonstrou ser eficaz para estruturar e analisar dados técnicos de patentes, mesmo com uma amostra reduzida. A implementação do pipeline completo — desde a extração e transformação de arquivos XML até a disponibilização via API RESTful — permitiu validar a viabilidade técnica da proposta, além de comprovar a capacidade do modelo dimensional em suportar análises OLAP e alimentar algoritmos de Inteligência Artificial, como redes neurais para previsão de termos.

Para trabalhos futuros, propõe-se a expansão da base de dados com maior volume e diversidade de patentes, o que poderá enriquecer as análises, aumentar a confiabilidade dos modelos preditivos e viabilizar estudos mais robustos em prospecção tecnológica e inteligência competitiva. Complementarmente, pretende-se integrar novas fontes de dados externos, como artigos científicos e notícias, permitindo abordagens multimodais de análise e ampliando o potencial de inovação da plataforma desenvolvida.

Referências

[1] WIPO – World Intellectual Property Organization. World Intellectual Property Indicators 2023. Disponível em: https://www.wipo.int/publications/en/details.jsp?id=4656

[2] R. R. Nogueira e A. L. G. Gonçalves, Aspectos técnicos e tecnológicos relacionados à análise de patentes: uma revisão integrativa. Anais do Simpósio de Pesquisa e Extensão (SIMPEX), SENAI, 2025.

[3] S. Peng et al., Automated patent analysis: NLP for structured and unstructured patent data, Proceedings of the PatentSemTech Workshop at CLEF 2021, CEUR Workshop Proceedings, 2020.

[4] R. Krestel, T. I. Bajwa, R. Drewes, Automated patent analysis: NLP for structured and unstructured patent data, CLEF 2021.

[5] R. Kimball, M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd ed., Wiley, 2013.

[6] W. H. Inmon, Building the Data Warehouse, 4th ed., Wiley, 2005.

[7] P. Gupta et al., Automatic Text Summarization Using Sequence-to-Sequence Model and Recurrent Neural Network, Computación y Sistemas, vol. 28, n. 4, 2024.

[8] W. M. F. Zaini, D. T. C. Lai, R. Chong Lim, Identifying patent classification codes associated with specific search keywords using machine learning, World Patent Information, vol. 71, 2022.

[9] Y. Zhao et al., Keyword extraction techniques for patent text analysis, 2019.

[10] S. Björkqvist, J. Kallio, Building a graph-based patent search engine, Proc. of the 46th Int. ACM SIGIR Conf., 2023, pp. 3300–3304.

[11] Y. Zhou et al., Forecasting emerging technologies with deep learning and data augmentation: convergence emerging technologies vs non-convergence emerging technologies, Proc. of 6th Int. Conf. on Future-Oriented Technology Analysis (FTA), 2019.

[12] S. M. González, T. C. Sakata, R. R. Nogueira, Newsminer: Enriched Multidimensional Corpus for Text-Based Applications, ICAISC 2020, Part II, Springer, 2020, pp. 231–242.

[13] R. M. Almeida, Ferramentas de Business Intelligence Open Source aplicadas à gestão do conhecimento em instituições de ensino superior, Revista GUAL, vol. 4, n. 2, pp. 145–164, 2011.

[14] P. Vyas, Demystifying Dimensional Modeling for Modern Data Warehousing, Journal of Computer Science and Technology Studies, vol. 7, n. 2, pp. 174–180, 2025.