

**UNIVERSIDADE DO ESTADO DE SANTA CATARINA – UDESC
CENTRO DE EDUCAÇÃO SUPERIOR DO ALTO VALE DO ITAJAÍ – CEAVI
ENGENHARIA DE SOFTWARE**

TOBIAS FELIPE KIEFER

**ARQUITETURA DE DATA WAREHOUSE AUTOMATIZADO PARA APOIO À
GESTÃO DO CONHECIMENTO EM BASES DE PATENTES**

IBIRAMA

2025

TOBIAS FELIPE KIEFER

**ARQUITETURA DE DATA WAREHOUSE AUTOMATIZADO PARA APOIO À
GESTÃO DO CONHECIMENTO EM BASES DE PATENTES**

Trabalho de conclusão apresentado ao curso de Engenharia de Software do Centro de Educação Superior do Alto Vale do Itajaí (CEAVI), da Universidade do Estado de Santa Catarina (UDESC), como requisito parcial para a obtenção do grau de bacharel em Engenharia de Software.

Orientador: Prof. MSc. Pedro Sidnei Zanchett
Coorientador: Prof. MSc. Rodrigo Ramos Nogueira

IBIRAMA

2025

TOBIAS FELIPE KIEFER

**ARQUITETURA DE DATA WAREHOUSE AUTOMATIZADO PARA APOIO À
GESTÃO DO CONHECIMENTO EM BASES DE PATENTES**

Trabalho de conclusão apresentado ao curso de Engenharia de Software do Centro de Educação Superior do Alto Vale do Itajaí (CEAVI), da Universidade do Estado de Santa Catarina (UDESC), como requisito parcial para a obtenção do grau de bacharel em Engenharia de Software.

Orientador: Prof. MSc. Pedro Sidnei Zanchett

Coorientador: Prof. MSc. Rodrigo Ramos Nogueira

BANCA EXAMINADORA:

Profº. Pedro Sidnei Zanchett, MSc.
UDESC

Membros:

Profº. Carlos Alberto Barth, TMSc.
UDESC

Profº. Geraldo Menegazzo Varela, MSc.
UDESC

Ibirama, 03 de julho de 2025

RESUMO

Com o crescimento exponencial de documentos técnicos e científicos, a análise automatizada de informações contidas em bases de patentes tornou-se essencial para subsidiar decisões estratégicas em contextos de inovação. Este trabalho apresenta o desenvolvimento de uma arquitetura automatizada de *Data Warehouse*, voltada à extração, organização e análise de dados textuais provenientes de arquivos XML do USPTO (*United States Patent and Trademark Office*). A solução proposta é composta por um pipeline de ETL desenvolvido em Python, que realiza a transformação e carga dos dados em um banco de dados PostgreSQL modelado segundo a abordagem dimensional. Adicionalmente, foi implementada uma API REST com consultas analíticas, permitindo a exploração dos dados por diferentes dimensões, como autor, país, tempo e termos técnicos. Um diferencial da proposta é a utilização de modelos de linguagem natural para extração automática de entidades técnicas, que são integradas ao modelo dimensional para enriquecer semanticamente as análises. Os resultados demonstraram a viabilidade técnica da solução, que se mostrou eficaz para consultas exploratórias, detecção de tendências e apoio à gestão do conhecimento tecnológico. A arquitetura desenvolvida contribui para a democratização do acesso a informações tecnológicas estruturadas, oferecendo uma base sólida para aplicações em inteligência competitiva e prospecção de inovação.

Palavras-chave: *Data Warehouse*; ETL Automatizado; Patentes; Análise Textual; Inteligência Artificial.

ABSTRACT

With the exponential growth of technical and scientific documents, the automated analysis of information contained in patent databases has become essential for supporting strategic decision-making in innovation contexts. This work presents the development of an automated *Data Warehouse* architecture designed for extracting, organizing, and analyzing textual data from XML files provided by the USPTO (United States Patent and Trademark Office). The proposed solution includes an ETL pipeline developed in Python, which transforms and loads the data into a PostgreSQL database modeled using the dimensional approach. Additionally, a RESTful API was implemented to enable OLAP-based analytical queries, allowing data exploration through multiple dimensions such as author, country, time, and technical terms. A key feature of this architecture is the integration of large language models (LLMs) for the automatic extraction of technical entities, which are incorporated into the dimensional model to semantically enrich the analysis. The results confirmed the technical feasibility of the solution, which proved effective for exploratory queries, trend detection, and knowledge management support. The developed architecture contributes to democratizing access to structured technological information, offering a solid foundation for applications in competitive intelligence and innovation forecasting.

Keywords: Data Warehouse; Automated ETL; Patents; Text Analysis; Artificial Intelligence.

LISTA DE ABREVIATURAS E SIGLAS

PLN	Processamento de Linguagem Natural
ETL	Extract, Transform, Load
IA	Inteligência Artificial
API	Application Programming Interface
OLAP	Online Analytical Processing
USPTO	United States Patent and Trademark Office
PCT	Patent Cooperation Treaty)
INPI	Instituto Nacional da Propriedade Industrial
BI	Business Intelligence
XML	Extensible Markup Language
NER	Named Entity Recognition
LLMs	Large Language Models

SUMÁRIO

1	INTRODUÇÃO	8
1.1	PROBLEMA	10
1.2	OBJETIVOS	11
1.2.1	Objetivo Geral	11
1.2.2	Objetivos Específicos	11
1.3	JUSTIFICATIVA	11
1.4	METODOLOGIA	12
2	FUNDAMENTAÇÃO TEÓRICA	14
2.1	PATENTES COMO FONTE DE CONHECIMENTO TECNOLÓGICO . . .	14
2.2	ARQUITETURA DE <i>DATA WAREHOUSE</i>	15
2.2.1	Modelagem Dimensional e Modelo Estrela	16
2.3	PROCESSOS DE EXTRAÇÃO, TRANSFORMAÇÃO E CARGA DE DADOS	17
2.3.1	Extração	18
2.3.2	Transformação	18
2.3.3	Carga	18
2.3.4	<i>Stopwords</i> e Redução de Ruído Textual	19
2.3.5	Extração de Entidades Nomeadas e Normalização	19
2.4	MODELOS DE IA PARA ANÁLISE DE DADOS TEXTUAIS	20
2.4.1	<i>Word Embeddings</i> e Representações Semânticas	20
2.4.1.1	<i>Vantagens dos Word Embeddings</i>	21
2.4.1.2	<i>Desafios e Limitações dos Word Embeddings</i>	21
2.5	TRABALHOS CORRELATOS	22
2.5.1	Modelagem e Infraestrutura de Dados	23
2.5.2	ETL Distribuído e Automação	23
2.5.3	Análise Semântica e Textual de Patentes	23
2.5.4	Técnicas Avançadas e Tarefas Específicas	23
2.5.5	Síntese e Contribuição do Presente Trabalho	24
3	DESENVOLVIMENTO	25
3.1	VISÃO GERAL DA ARQUITETURA	25
3.2	MODELAGEM DO <i>DATA WAREHOUSE</i>	26
3.3	DESENVOLVIMENTO DO <i>PIPELINE</i> ETL	28
3.3.1	Extração	28
3.3.2	Transformação	31
3.3.3	Carga	32
3.3.4	Execução Local e em Nuvem via Supabase	33

3.4	AUTOMAÇÃO COM DOCKER	34
3.5	API REST PARA CONSULTA DE DADOS	37
3.5.1	Estrutura da API	37
3.5.2	Principais Endpoints da API REST	37
3.5.3	Consultas OLAP na API	38
3.5.4	Parâmetros Dinâmicos	39
3.5.5	Benefícios da API REST	39
3.6	EXECUÇÃO DO PROJETO	39
3.6.1	Pré-requisitos	39
3.6.2	Clonar o Repositório GitHub	39
3.6.3	Baixar o Arquivo XML da USPTO	39
3.6.4	Executar o Sistema com Docker	40
4	RESULTADOS E DISCUSSÕES	41
4.1	EXECUÇÃO DO <i>PIPELINE</i> ETL	41
4.2	QUALIDADE E CONSISTÊNCIA DOS DADOS CARREGADOS	41
4.3	EXTRAÇÃO DE ENTIDADES TÉCNICAS COM APOIO DE LLMS	42
4.4	VALIDAÇÃO DO MODELO MULTIDIMENSIONAL E CONSULTAS OLAP	42
4.4.1	Integração com o Modelo Dimensional	43
4.4.2	Parâmetros Experimentais	44
4.4.3	Experimento 1: Análise Temporal de Termos Técnicos por Ano	45
4.4.3.1	<i>Explorando Dimensões – Categoria</i>	47
4.4.3.2	<i>Explorando Dimensões – Autores</i>	48
4.4.3.3	<i>Explorando Dimensões – Países</i>	48
4.4.4	Associação entre Termos Técnicos	49
4.4.5	Associação de Termos Técnicos por Subclasse Tecnológica	49
4.4.6	Análise Temporal de Tendências Tecnológicas com Modelos Preditivos	51
4.4.7	Agrupamento de Patentes com K-Means e Análise de Clusters Tecnológicos	54
4.5	CONSUMO DE DADOS VIA API REST	57
4.6	ANÁLISE CRÍTICA DA ARQUITETURA	58
4.7	DISCUSSÃO DOS RESULTADOS	59
4.8	LIMITAÇÕES E PERSPECTIVAS FUTURAS	59
5	CONCLUSÃO	62
	REFERÊNCIAS	63

1 INTRODUÇÃO

Em um cenário de intensa transformação digital, a inovação tecnológica consolidou-se como um dos pilares fundamentais para a competitividade entre empresas, universidades e instituições de pesquisa. Nesse contexto, o acesso a informações técnicas confiáveis e atualizadas torna-se estratégico. As bases de dados de patentes destacam-se como fontes ricas, organizadas e estruturadas de conhecimento técnico, oferecendo uma visão aprofundada sobre o estado da arte em diversas áreas do saber.

Esses repositórios não apenas documentam invenções e descobertas, como também trazem metadados relevantes tais como autores, datas de registro, classificações tecnológicas e descrições técnicas completas que podem ser explorados com fins analíticos e preditivos. Além de servirem como mecanismo de proteção legal para as inovações, as patentes desempenham um papel crucial na difusão do conhecimento tecnológico. Isso ocorre porque, ao garantir direitos de exclusividade ao inventor, o sistema de patentes exige a divulgação detalhada da tecnologia desenvolvida, contribuindo diretamente para o avanço coletivo da ciência e da engenharia.

Além de seu papel técnico e jurídico, as patentes possuem um impacto significativo no desenvolvimento econômico e social. Ao documentar avanços tecnológicos de forma estruturada e pública, elas promovem a disseminação do conhecimento e incentivam a inovação aberta. A análise dessas bases, portanto, não é apenas um instrumento de vantagem competitiva para empresas e instituições, mas também um meio de apoiar políticas públicas, fomentar o empreendedorismo e orientar investimentos em pesquisa e desenvolvimento com base em evidências concretas de evolução tecnológica.

Com o avanço das tecnologias de análise de dados, documentos de patentes passaram a ser amplamente utilizados como insumo para atividades de inteligência tecnológica. Técnicas de Inteligência Artificial (IA), como aprendizado de máquina e grafos de conhecimento, permitem identificar padrões ocultos, antecipar movimentos tecnológicos e mapear tendências emergentes. A análise automatizada de grandes volumes de dados patentários tem se mostrado eficiente para a descoberta de relações semânticas entre tecnologias e previsão de caminhos inovadores futuros (Peng et al., 2020).

Apesar de sua relevância estratégica, extrair valor informacional dessas bases é um desafio não trivial, devido ao grande volume de dados, à complexidade dos formatos frequentemente disponibilizados em arquivos XML (*Extensible Markup Language*) e à necessidade de técnicas avançadas para organização, análise e visualização. Tradicionalmente, essas bases são subutilizadas em projetos de ciência de dados e inteligência competitiva, em parte pela ausência de ferramentas acessíveis que permitam automatizar o tratamento e a análise dessas informações.

Os arquivos de patentes do *United States Patent and Trademark Office* (USPTO) são fornecidos em pacotes semanais em formato XML, com estrutura hierárquica densa e muitas vezes inconsistente entre versões. Isso torna o processo de extração automatizada mais complexo e suscetível a erros de interpretação, perda de dados relevantes e dificuldades na integração com

bancos analíticos e ferramentas de *Business Intelligence* (BI) (Krestel et al., 2021).

Um *Data Warehouse* é uma estrutura projetada para integrar dados provenientes de múltiplas fontes, mantendo características como integração, orientação por assunto, não-volatilidade e variação temporal (Kimball; Ross, 2013). Essas propriedades o tornam especialmente adequado para lidar com grandes volumes de dados históricos e heterogêneos, como é o caso das bases de patentes disponibilizadas semanalmente pelo USPTO. A cada semana, são liberados aproximadamente 1 GB de dados no formato XML, contendo milhares de registros técnicos altamente detalhados. Esses arquivos apresentam desafios significativos de coleta, extração e padronização, dada sua estrutura complexa, seu alto volume e sua recorrência temporal, características que a literatura aponta como inerentes aos desafios de Big Data (Xu et al., 2016). Diante disso, este trabalho propõe o desenvolvimento de uma arquitetura de *Data Warehouse* automatizado que realize a extração, transformação, carga (ETL) e análise desses dados. O objetivo de uma ferramenta assim é apoiar a gestão do conhecimento e a tomada de decisão estratégica por meio da consolidação e organização dessas informações. Em ambientes de inovação tecnológica, onde o volume de novos registros cresce continuamente, contar com uma base estruturada permite identificar tendências emergentes, mapear áreas de desenvolvimento e reduzir a assimetria de informações entre organizações. Conforme destacado por Inmon (2005), a qualidade e organização dos dados em um *Data Warehouse* impactam diretamente a capacidade de gerar conhecimento útil para decisões orientadas por dados, que poderão ser consumidas nos mais diversos ambientes, sejam eles acadêmicos, industriais e governamentais.

A arquitetura desenvolvida é composta por três camadas principais, cada uma projetada para lidar com os desafios específicos do processamento de dados do USPTO, tais como o grande volume de arquivos disponibilizados semanalmente em formato XML, sua complexidade estrutural e a necessidade de análises históricas.

A primeira camada consiste em um *pipeline* automatizado de ETL, responsável por ler, interpretar e converter os dados brutos dos arquivos XML da USPTO em estruturas tabulares consistentes. Esse processo envolve a remoção de ruídos e inconsistências no formato dos arquivos, a padronização de nomes, datas e campos textuais, e a extração de metadados técnicos como títulos de invenções, inventores, países de origem, resumos e datas de publicação. O resultado é uma base de dados limpa e estruturada, preparada para análise posterior.

A segunda camada é um *Data Warehouse* modelado de forma multidimensional, com a utilização de uma tabela fato central, responsável por guardar a ocorrência de palavras em cada patente, conectada a diversas tabelas de dimensão, como ano de publicação, setor tecnológico, autor, país e palavras-chave. Essa modelagem permite armazenar grandes volumes de dados históricos organizados temporalmente, favorecendo a análise evolutiva de tecnologias ao longo do tempo, o cruzamento de variáveis estratégicas e a construção de indicadores relevantes para apoio à tomada de decisão em contextos de inovação e inteligência tecnológica.

Por fim, a terceira camada da arquitetura corresponde à interface de acesso aos dados, que oferece suporte a consultas analíticas avançadas por meio de operações OLAP (Online

Analytical Processing). Esse tipo de processamento permite explorar os dados de forma multidimensional, viabilizando análises por diferentes perspectivas como tempo, país, autor ou categoria tecnológica. Além disso, essa camada contempla mecanismos de visualização interativa para facilitar a interpretação dos resultados e apoiar a tomada de decisão. Também está preparada para integração com modelos de IA voltados ao processamento de linguagem natural, viabilizando análises semânticas mais sofisticadas, como extração automatizada de palavras-chave, detecção de entidades técnicas relevantes e identificação de tendências emergentes nos conteúdos das patentes.

Com isso, o trabalho busca não apenas construir uma infraestrutura técnica robusta, mas também propor uma abordagem prática e replicável para projetos que envolvam grandes volumes de dados textuais especializados, como é o caso das patentes.

1.1 PROBLEMA

Apesar de sua relevância estratégica, as bases de patentes, como as disponibilizadas pelo USPTO, apresentam desafios significativos para sua utilização em projetos de inteligência competitiva e análise tecnológica. Essas bases são disponibilizadas semanalmente em grandes volumes de dados e estruturadas em arquivos no formato XML semiestruturado, o que exige técnicas especializadas para extração e interpretação das informações.

A estrutura dos documentos XML da USPTO é altamente detalhada, contendo múltiplos níveis de aninhamento e campos técnicos, como resumos, classificações, inventores, requerentes e datas de publicação. No entanto, sua complexidade torna difícil a aplicação direta de métodos tradicionais de análise de dados, além de demandar grande esforço computacional e conhecimento técnico para o pré-processamento. A ausência de uma estrutura relacional clara dificulta a padronização dos dados e sua integração com ferramentas analíticas convencionais.

Estudos recentes apontam que essas limitações contribuem para a subutilização do potencial estratégico dos dados de patentes, sobretudo em iniciativas que demandam análise em larga escala, como mineração de texto, detecção de tendências e apoio à formulação de políticas de inovação (González; Sakata; Nogueira, 2020). O caráter temporal dos dados, com novas patentes sendo registradas constantemente, demanda soluções que possibilitem a atualização contínua e eficiente dos dados armazenados.

Dado essa problemática, torna-se evidente a necessidade de arquiteturas que apoiem não apenas a coleta automatizada desses arquivos, mas também sua transformação em estruturas compatíveis com consultas analíticas. A construção de um *Data Warehouse*, com etapas bem definidas de ETL e modelagem multidimensional, se apresenta como uma estratégia adequada para enfrentar esses desafios, proporcionando maior acessibilidade e valor estratégico aos dados de patentes.

1.2 OBJETIVOS

Esta seção apresenta os objetivos geral e específicos do trabalho.

1.2.1 Objetivo Geral

O objetivo principal deste trabalho é desenvolver e avaliar uma arquitetura de *Data Warehouse* automatizado para a coleta, organização e análise de grandes volumes de dados de patentes, visando apoiar a gestão do conhecimento e a tomada de decisão estratégica.

1.2.2 Objetivos Específicos

Para atingir o principal objetivo deste trabalho, definem-se os seguintes objetivos específicos:

- Modelar e implementar um *Data Warehouse* multidimensional para estruturar dados coletados de bases públicas de patentes, como a USPTO;
- Desenvolver e validar um *pipeline* automatizado de ETL para garantir a atualização periódica e confiável dos dados;
- Implementar uma *Application Programming Interface* (API) para disponibilização e consumo dos dados armazenados no *Data Warehouse*, facilitando a integração com aplicações externas e sistemas de apoio à decisão;
- Realizar análises exploratórias e descritivas sobre os dados armazenados, utilizando consultas analíticas para identificar tendências tecnológicas, áreas emergentes e padrões de inovação;
- Aplicar técnicas de processamento e análise textual para enriquecer as informações contidas nas bases de patentes, apoiando a gestão do conhecimento organizacional;
- Propor recomendações sobre o uso de *Data Warehousing* como ferramenta estratégica na gestão do conhecimento e apoio à inovação, especialmente no contexto de análise de patentes.

1.3 JUSTIFICATIVA

As patentes são mais do que simples registros legais de propriedade intelectual: são fontes valiosas de conhecimento técnico, tendências de inovação e informação estratégica. Em países desenvolvidos, bancos de dados de patentes já são amplamente explorados para fins de prospecção tecnológica, mapeamento competitivo e apoio à pesquisa e desenvolvimento (Peng et al., 2020). No entanto, o acesso e o uso eficaz dessas bases ainda enfrentam grandes barreiras técnicas, especialmente em ambientes com recursos computacionais ou humanos limitados.

Boa parte dessas dificuldades reside na natureza semiestruturada dos arquivos XML, que exigem conhecimento técnico especializado para extração e transformação dos dados. Além disso, muitas vezes não há integração direta entre essas fontes e ferramentas analíticas que

suportem consultas multidimensionais ou integração com modelos de IA textual. Isso representa um gargalo para pesquisadores, analistas e gestores interessados em utilizar essas informações de forma estratégica e automatizada.

A proposta de um *Data Warehouse* automatizado surge como resposta eficaz a esse cenário. Estruturar os dados de patentes em uma arquitetura analítica baseada em modelos dimensionais permite a realização de análises temporais, comparativas e segmentadas (Kimball; Ross, 2013). Por sua vez, a automação do *pipeline* ETL reduz o custo operacional de atualização dos dados e melhora a confiabilidade do processo, possibilitando ciclos contínuos de ingestão e análise.

A integração desses dados com técnicas de IA e PLN, como sumarização automática, classificação por área tecnológica e extração de palavras-chave, permite escalar a extração de valor informacional. Isso pode transformar o *Data Warehouse* em um verdadeiro sistema de apoio à decisão, aplicável tanto em contextos acadêmicos quanto industriais e institucionais.

Nesse sentido, este trabalho contribui de maneira prática e relevante para a automação da análise de patentes, promovendo o uso inteligente e inovador de dados abertos. A proposta também oferece um modelo replicável, que pode ser adaptado para outras fontes de dados textuais e domínios do conhecimento.

1.4 METODOLOGIA

Este trabalho caracteriza-se como uma pesquisa tecnológica aplicada e exploratória, conforme a classificação de Junior et al. (JUNIOR et al., 2014), voltada ao desenvolvimento e validação de uma arquitetura automatizada de *Data Warehouse* com foco na coleta, estruturação e análise de dados extraídos de documentos de patentes. O processo metodológico foi organizado em etapas sequenciais que compreenderam a modelagem multidimensional dos dados, a implementação de um *pipeline* de ETL automatizado, o desenvolvimento de uma API de acesso e a integração com ferramentas analíticas e modelos de IA.

A modelagem do *Data Warehouse* foi fundamentada na análise estrutural dos arquivos XML disponibilizados pelo USPTO. Esses arquivos apresentam dados complexos, distribuídos em diferentes camadas, contendo informações como resumos técnicos, classificações, datas, autores e localizações geográficas. Para tratar essa diversidade, foi adotado o esquema estrela, cuja tabela fato central (*fact_patents*) registra as ocorrências de palavras nos resumos de patentes. As dimensões incluem autores (*dim_authors*), países (*dim_countries*), datas (*dim_date*), patentes (*dim_patents*) e palavras (*dim_words*). Essa estrutura viabiliza consultas analíticas por tempo, autoria, nacionalidade e termos técnicos, alinhando-se a recomendações de Kimball e Ross (2013) para projetos de armazenagem analítica de dados com recorte temporal e semiestruturado.

A extração e transformação dos dados foram conduzidas por um *pipeline* ETL desenvolvido em Python, que operacionaliza a coleta automatizada dos arquivos XML da USPTO, com

foco na remoção de metadados repetitivos, padronização de datas e normalização de nomes. Na fase de transformação, aplicaram-se técnicas de tokenização e limpeza textual, convertendo o conteúdo técnico em formato estruturado. A etapa de carga organiza os dados em uma camada intermediária (*staging*) e os distribui posteriormente entre a tabela fato e as tabelas de dimensão do *Data Warehouse*. Essa abordagem segue estratégias bem-sucedidas já documentadas na literatura para o tratamento de grandes volumes de dados semiestruturados González, Sakata e Nogueira (2020).

A arquitetura proposta inclui ainda uma API RESTful construída com o framework FastAPI, permitindo acesso externo aos dados consolidados. A API oferece suporte a filtros por país, autor, período e palavras-chave, além de integrar visualizações interativas e alimentar painéis analíticos com dados oriundos do *Data Warehouse*.

Na fase de análise, foram realizadas consultas analíticas em SQL com suporte a operações OLAP, visando compreender padrões temporais, associações semânticas e distribuição geográfica de depósitos de patentes. As consultas facilitaram a interpretação do acervo textual, fornecendo insumos para a etapa complementar de aplicação de técnicas de IA.

Foi realizado a integração de modelos de IA treinados para tarefas de classificação automática de áreas tecnológicas, geração de resumos técnicos e avaliação de desempenho computacional. Os modelos foram configurados para consumir diretamente os dados organizados na camada de *Data Warehouse*, com o objetivo de potencializar a exploração semântica e analítica dos dados.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os fundamentos teóricos que sustentam o desenvolvimento da arquitetura proposta neste trabalho. A revisão da literatura foi organizada em quatro tópicos principais: (i) patentes como fonte de conhecimento tecnológico, (ii) arquitetura de *Data Warehouse* e modelagem dimensional, (iii) processos de Extração, Transformação e Carga (ETL), e (iv) técnicas de IA aplicadas à análise de dados textuais.

Cada seção discute conceitos fundamentais e abordagens existentes na literatura, que servem de base para as decisões técnicas e metodológicas adotadas ao longo deste projeto.

2.1 PATENTES COMO FONTE DE CONHECIMENTO TECNOLÓGICO

Uma patente é um título de propriedade temporária concedido pelo Estado a um inventor ou titular, garantindo o direito de exclusividade sobre a exploração de uma invenção por um período determinado, geralmente de 20 anos. Em contrapartida, o inventor deve tornar pública uma descrição completa da invenção, permitindo que outras pessoas compreendam seu funcionamento. Trata-se, portanto, de um mecanismo legal que estimula a inovação ao recompensar a criatividade técnica e ao mesmo tempo disseminar conhecimento.

As patentes servem a múltiplos propósitos: proteger invenções contra uso indevido por terceiros, atrair investimentos ao garantir exclusividade de mercado, e fomentar a troca de informações técnicas e científicas. Entre os principais tipos de patente estão: patentes de invenção e modelos de utilidade, variando conforme o grau de inovação e complexidade, de acordo com Instituto Nacional da Propriedade Industrial (INPI) (2025).

Os elementos principais de uma patente incluem:

- Título da invenção: denominação clara e sucinta do objeto protegido;
- Resumo: descrição breve do conteúdo técnico;
- Relatório descritivo: detalhamento do funcionamento da invenção;
- Reivindicações: delimitam juridicamente o escopo da proteção;
- Desenhos: ilustrações que complementam a compreensão da invenção;
- Informações bibliográficas: como número do pedido, data de depósito, país de origem, nome dos inventores e classificação técnica.

Patentes representam, portanto, uma das formas mais estruturadas e acessíveis de conhecimento tecnológico disponível publicamente. Ao registrar uma invenção, o autor ou organização responsável divulga, em detalhes, o funcionamento, aplicação e originalidade da tecnologia desenvolvida. Essa divulgação é parte essencial do contrato de concessão de patentes, que garante o direito de exclusividade em troca da publicação do conhecimento técnico.

Além disso, patentes possuem um formato padronizado internacionalmente, seguindo convenções como o PCT (*Patent Cooperation Treaty*), o que favorece sua análise automatizada (World Intellectual Property Organization (WIPO), 2025). Bases como o USPTO, o *Espacenet* da *European Patent Office* e o Instituto Nacional da Propriedade Industrial (INPI) (2025) oferecem

milhares de registros atualizados com metadados como título da invenção, resumo, área técnica, inventores, datas de depósito, classificação internacional, entre outros.

Essas bases, contudo, apresentam desafios relevantes para seu aproveitamento analítico. A estrutura XML dos documentos, embora organizada, exige ferramentas específicas para leitura e extração dos dados. Além disso, a natureza textual e descritiva dos resumos e descrições das patentes requer o uso de técnicas de PLN para identificar padrões, conceitos-chave e relações semânticas.

Sob a perspectiva da gestão do conhecimento, os dados provenientes de patentes constituem um ativo informacional de grande valor estratégico. No entanto, para que esse potencial seja plenamente explorado, é necessário que tais dados estejam organizados de forma estruturada e acessível. Nesse contexto, o uso de um *Data Warehouse* se mostra essencial, ao permitir o armazenamento histórico, a categorização por temas técnicos e a análise temporal de registros, transformando coleções massivas de documentos em uma base sólida para apoio à inovação, pesquisa e desenvolvimento.

A integração dessas bases patentárias em arquiteturas de BI e *Data Warehousing* viabiliza não apenas a sistematização e a consulta eficiente dos dados, como também abre espaço para a aplicação de técnicas analíticas avançadas e de IA. Com isso, torna-se possível extrair padrões, detectar tendências emergentes e gerar *insights* relevantes a partir de fontes abertas, amplificando o valor do conhecimento tecnológico contido nas patentes.

2.2 ARQUITETURA DE DATA WAREHOUSE

Um *Data Warehouse* é uma estrutura de dados voltada para suporte à decisão, cuja função principal é armazenar dados históricos, integrados e organizados de forma a facilitar consultas analíticas complexas e geração de relatórios gerenciais. Ao contrário dos sistemas transacionais, cujo foco é o processamento de operações do dia a dia, o *Data Warehouse* tem como prioridade a análise de grandes volumes de dados ao longo do tempo, sendo um pilar fundamental em projetos de BI e gestão do conhecimento.

Segundo Inmon (2005), considerado um dos precursores do conceito, um *Data Warehouse* é uma "coleção de dados orientada por assunto, integrada, não volátil e variante no tempo, organizada para apoiar processos de tomada de decisão". Essa definição destaca quatro características essenciais (Inmon, 2005):

- Orientado por assunto: os dados são organizados em torno de temas de interesse, como clientes, produtos, regiões ou, neste caso, patentes;
- Integrado: dados provenientes de diversas fontes são padronizados e consolidados;
- Não volátil: uma vez carregados, os dados não sofrem alterações frequentes;
- Variável no tempo: os dados mantêm seu histórico e são organizados em função de datas e períodos.

Para representar os dados em um *Data Warehouse*, é comum utilizar a modelagem

dimensional, baseada em dois tipos principais de tabelas:

- Tabelas de dimensão: armazenam atributos descritivos que caracterizam os dados, como autor da patente, país de origem, data do depósito, entre outros;
- Tabelas fato: centralizam os eventos quantitativos e relacionam-se com as dimensões por meio de chaves estrangeiras. No presente trabalho, a tabela fato armazena informações sobre a ocorrência de palavras nos resumos das patentes, ligando-as às demais dimensões.

A modelagem do tipo estrela (*star schema*) é amplamente utilizada por sua simplicidade e desempenho em consultas analíticas. Segundo Kimball e Ross (2013), essa abordagem favorece a performance das consultas OLAP e torna o modelo acessível para usuários de negócios e analistas de dados. A simplicidade e o desempenho em consultas analíticas são as principais razões para a ampla utilização da modelagem estrela.

No contexto deste trabalho, a arquitetura do *Data Warehouse* foi planejada para integrar dados extraídos de arquivos XML da base USPTO, estruturando-os em um modelo analítico capaz de responder a perguntas como:

- Quais os inventores mais produtivos por país?
- Quais palavras-chave se destacam em pedidos recentes?
- Quais áreas tecnológicas mostram crescimento ao longo do tempo?

Além disso, a arquitetura permite a extensão futura para análise textual e aplicação de algoritmos de IA, conforme discutido nas seções seguintes.

2.2.1 Modelagem Dimensional e Modelo Estrela

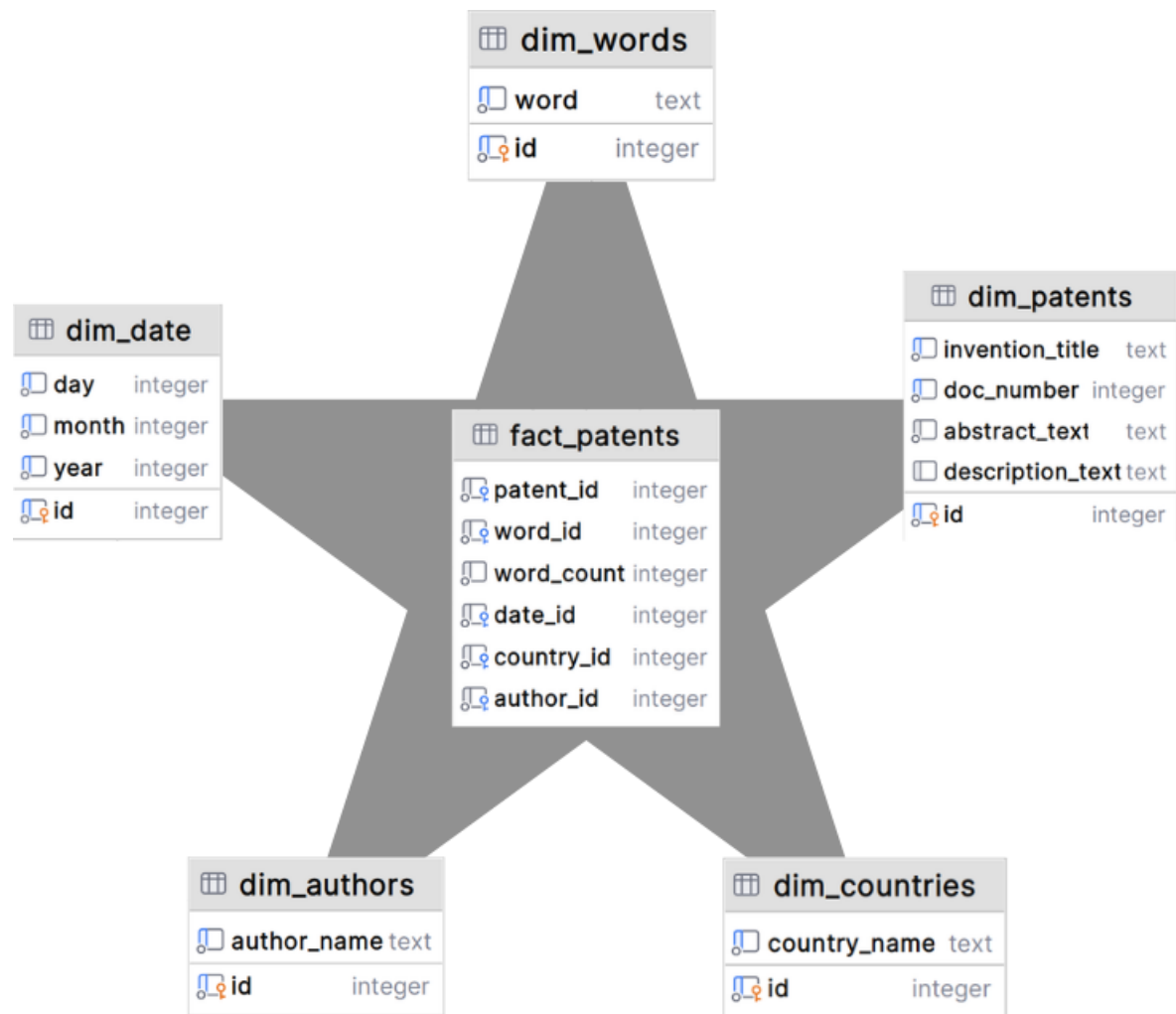
A modelagem dimensional é uma técnica de estruturação de dados voltada para ambientes de apoio à decisão, especialmente sistemas de BI e *Data Warehouses*. Diferente dos modelos normalizados tradicionais utilizados em bancos de dados transacionais (*OLTP*), a modelagem dimensional prioriza a simplicidade, a legibilidade e a performance de leitura, favorecendo consultas analíticas complexas.

Um dos principais formatos utilizados nessa abordagem é o modelo estrela (*star schema*), caracterizado por uma tabela fato central conectada a várias tabelas de dimensão. A tabela fato armazena medidas quantitativas como valores, contagens ou métricas de ocorrência que podem ser analisadas sob diferentes perspectivas tempo, local, autor, etc. As dimensões, por sua vez, armazenam atributos descritivos que fornecem o contexto da análise.

Segundo Nogueira (2020), o modelo estrela é a estrutura mais simples e eficiente para ambientes de BI que exigem respostas rápidas a grandes volumes de dados, justamente por reduzir a complexidade de interligações entre as tabelas e permitir fácil navegação pelas dimensões.

No contexto deste trabalho, o modelo estrela foi adotado para armazenar a contagem de palavras extraídas de resumos de patentes. A tabela fato registra a quantidade de ocorrências de palavras em resumos, enquanto as dimensões representam autores, datas, países, palavras e metadados da patente. Essa estrutura facilita operações OLAP, como análise temporal, ranking de termos técnicos e descoberta de padrões semânticos entre palavras, autores e países.

Figura 1 – Modelo estrela da solução proposta



Fonte: Elaborado pelo autor, 2025.

A Figura 1 ilustra a modelagem dimensional aplicada neste projeto. No centro, encontra-se a tabela fato **fact_patents**, responsável por registrar cada ocorrência de palavra em uma patente, ligada por chaves estrangeiras às tabelas de dimensão **dim_authors**, **dim_countries**, **dim_date**, **dim_words** e **dim_patents**. Esse arranjo permite múltiplas perspectivas de análise sobre os dados textuais, como filtragens por país, autor ou período, seguindo a lógica clássica do modelo estrela, onde o foco principal será sempre a tabela fato.

2.3 PROCESSOS DE EXTRAÇÃO, TRANSFORMAÇÃO E CARGA DE DADOS

O processo de ETL é uma das etapas fundamentais na construção e manutenção de um *Data Warehouse*. Sua função é extrair dados brutos de diferentes fontes, realizar transformações necessárias para padronização e qualidade da informação e, então, carregar os dados resultantes nas estruturas analíticas do *Data Warehouse*.

De acordo com Golfarelli e Rizzi (2009), o processo ETL pode ser dividido em três etapas principais:

2.3.1 Extração

Consiste em coletar dados de diferentes fontes, que podem ser bancos relacionais, planilhas, *APIs* ou arquivos semiestruturados, como XML e *JSON*. No presente trabalho, a fonte primária são os arquivos XML disponibilizados pelo USPTO, que contêm registros de pedidos de patentes com seus respectivos metadados e textos descritivos;

2.3.2 Transformação

Refere-se ao processo de tratamento e enriquecimento dos dados extraídos. Isso inclui limpeza de ruídos, correção de inconsistências, padronização de formatos, derivação de campos e aplicação de regras de negócio. Neste projeto, as transformações realizadas envolvem a padronização dos nomes de autores, a conversão de datas para o formato ISO, a tokenização dos resumos e a extração de palavras para posterior análise semântica. Também é aplicada a extração das chamadas *stopwords*, que são palavras com baixo valor semântico (como “e”, “de”, “para”), as quais podem ser removidas para refinar a qualidade da análise. Esse tipo de filtragem contribui para reduzir o ruído e melhorar o desempenho de algoritmos de *Text Mining* e BI;

2.3.3 Carga

Etapas finais em que os dados transformados são inseridos na base de dados analítica. Neste trabalho, a carga é realizada inicialmente em uma *staging area*, utilizada como camada intermediária para validação e tratamento incremental. Em seguida, os dados são distribuídos nas tabelas de dimensões e na tabela fato do *Data Warehouse*.

A correta implementação do processo ETL é essencial para garantir a qualidade, consistência e confiabilidade das análises. Como destacado por Kimball e Ross (2013), o processo ETL frequentemente representa entre 60% e 80% do esforço total em projetos de *Data Warehouse*, sendo responsável por consolidar dados heterogêneos em uma base única e acessível para usuários analíticos.

No presente trabalho, o processo ETL foi desenvolvido integralmente em *Python*, utilizando bibliotecas como *ElementTree* para o parse de arquivos XML, *re* para expressão regular e *psycopg2* para conexão com o banco de dados PostgreSQL. A execução ocorre de forma automatizada via contêiner Docker, o que garante reprodutibilidade e portabilidade. Assim, novos arquivos podem ser processados com mínima intervenção manual, permitindo escalabilidade e redução de retrabalho.

Além das transformações estruturais, o sistema realiza também pré-processamentos linguísticos básicos nos resumos das patentes, como segmentação de sentenças, normalização de letras (conversão para minúsculas), remoção de pontuação, separação de palavras por espaços e remoção das *stopwords* para melhor análise.

2.3.4 *Stopwords* e Redução de Ruído Textual

Em tarefas de análise textual, como aquelas aplicadas aos resumos de patentes neste trabalho, é comum encontrar palavras que aparecem com alta frequência, mas que têm baixo valor informacional. Termos como “de”, “e”, “para”, “em”, “com” e “entre” são exemplos típicos do que se chama de *stopwords*, palavras de função gramatical que geralmente não contribuem de forma significativa para a representação semântica do conteúdo (Mihalcea; Tarau, 2004).

A presença dessas palavras pode distorcer os resultados em análises baseadas em frequência, como contagem de palavras, coocorrência e agrupamento temático. Por isso, a remoção de *stopwords* é uma prática comum em pipelines de *Text Mining* e PLN, ajudando a reduzir o ruído e focar nos termos mais relevantes do domínio em estudo (Turney, 2000).

Neste trabalho, propõe-se a remoção de *stopwords* durante a etapa de transformação dos dados textuais. Para isso, utiliza-se a lista padrão da língua inglesa fornecida pela biblioteca *nltk*, que inclui termos de alta frequência e baixo valor semântico. Essa filtragem contribui para uma análise mais precisa, especialmente em tarefas baseadas em frequência e associação semântica.

2.3.5 Extração de Entidades Nomeadas e Normalização

Além da remoção de *stopwords*, técnicas mais avançadas de pré-processamento podem ser aplicadas para enriquecer a análise textual em ambientes de apoio à decisão, como os *Data Warehouses*. Duas dessas técnicas são a extração de entidades nomeadas (*Named Entity Recognition* NER) e a normalização textual, ambas fundamentais para aumentar a precisão e a utilidade das análises realizadas sobre grandes volumes de texto, como os contidos em documentos de patentes.

O NER é uma tarefa da área de PLN que visa identificar automaticamente elementos informacionais específicos em textos, tais como nomes de organizações, unidades de medida, tecnologias ou pessoas. De acordo com Nadeau e Sekine (2007), o NER é essencial para transformar texto não estruturado em dados úteis e semânticos, permitindo, por exemplo, que sistemas automatizados filtrem, agrupem ou conectem documentos com base nas entidades extraídas. Em contextos como o de patentes, essa técnica viabiliza a identificação de termos técnicos críticos que frequentemente aparecem em diferentes formas ou posições estruturais nos documentos.

Complementarmente, a normalização textual tem o objetivo de reduzir a variabilidade linguística dos termos extraídos, facilitando análises agregadas. Isso inclui operações como lematização (redução de palavras à sua forma canônica), padronização de siglas e correção de variações morfológicas. Essa padronização é particularmente importante em bases técnicas, pois termos como “bateria” e “baterias”, ou “análise” e “análises”, devem ser tratados de forma unificada para evitar distorções em análises estatísticas e semânticas. Segundo Tikk et al. (2010), a normalização melhora significativamente o desempenho de tarefas baseadas em extração de

relações e padrões de ocorrência em domínios com vocabulário especializado.

Como discutido por Inmon (2005), a qualidade dos dados carregados em um *Data Warehouse* afeta diretamente a qualidade da informação gerada para suporte à decisão e técnicas como NER e normalização desempenham papel estratégico nesse processo.

Neste trabalho, as técnicas de extração de entidades nomeadas e normalização textual foram parcialmente implementadas com o uso de modelos de linguagem de última geração (LLMs), os quais analisaram os títulos e resumos das patentes para extrair termos técnicos relevantes. Esses termos, após processo de normalização (com padronização textual e filtragem semântica), foram armazenados na coluna `generated_terms` da tabela `staging_patents` e integrados às tabelas `dim_words` e `fact_patents`, o que viabiliza análises quantitativas e temporais baseadas em conceitos técnicos extraídos automaticamente. Essa abordagem demonstra o potencial da aplicação de NER e normalização em pipelines de ETL voltados para ambientes analíticos, promovendo maior valor semântico aos dados estruturados no *Data Warehouse*.

2.4 MODELOS DE IA PARA ANÁLISE DE DADOS TEXTUAIS

A análise de dados textuais, especialmente em domínios técnicos como o de patentes, representa um desafio significativo devido à complexidade semântica, ao vocabulário especializado e ao grande volume de informações não estruturadas. Para enfrentar esses desafios, modelos de IA vêm sendo amplamente utilizados, com destaque para abordagens baseadas em PLN.

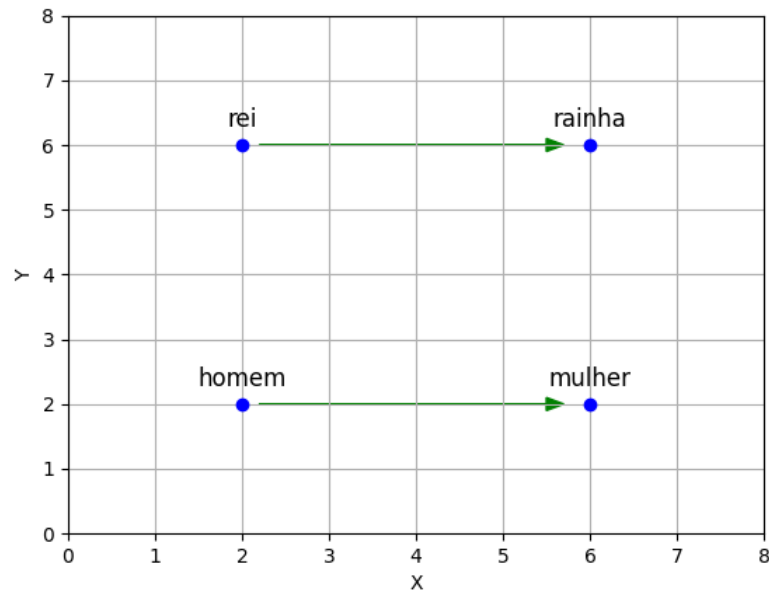
O PLN é um campo da IA que busca habilitar sistemas computacionais a compreender, interpretar e gerar linguagem humana. Dentro desse contexto, técnicas como representação vetorial de palavras (*Word Embeddings* e Representações Semânticas), classificação de texto, extração de palavras-chave e sumarização automática são amplamente empregadas para extrair conhecimento útil de textos longos e densos, como os encontrados em resumos e descrições de patentes.

2.4.1 *Word Embeddings* e Representações Semânticas

Modelos de representação como o Word2Vec (Mikolov et al., 2013), GloVe (Pennington; Socher; Manning, 2014) e FastText (Joulin et al., 2016) permitem transformar palavras em vetores numéricos densos que capturam relações semânticas e sintáticas entre os termos. Isso possibilita, por exemplo, identificar termos semelhantes, agrupamentos temáticos e relações latentes entre conceitos técnicos.

Modelos mais recentes como o BERT (Devlin et al., 2019) e o GPT (Radford et al., 2018) vão além ao considerar o contexto completo de cada palavra na frase, gerando embeddings contextuais que melhoram significativamente o desempenho em tarefas complexas de interpretação textual.

Figura 2 – Visualização de Embeddings para Relações Análogas de Gênero e Realeza



Fonte: Elaborado pelo autor, 2024.

2.4.1.1 Vantagens dos Word Embeddings

Conforme apresentado por Mikolov et al. (2013), os *word embeddings* têm várias vantagens sobre representações tradicionais de palavras, como o *one-hot encoding*. Uma das principais vantagens é a redução da dimensionalidade. Enquanto a representação *one-hot* requer que cada palavra seja representada por um vetor do tamanho do vocabulário, os *embeddings* representam palavras em vetores densos de uma dimensão muito menor, o que economiza espaço de memória e reduz a complexidade computacional.

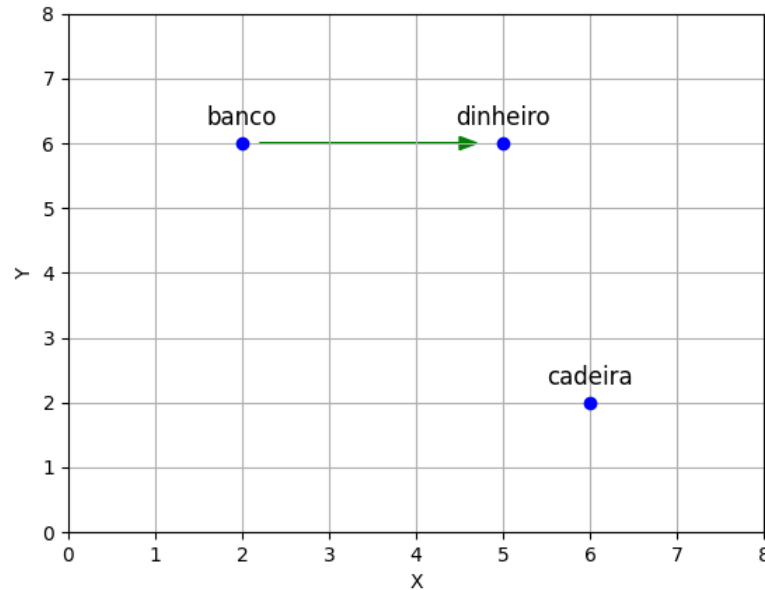
Além disso, os *embeddings* capturam relações semânticas e sintáticas de forma eficiente. Palavras com significados semelhantes são mapeadas para pontos próximos no espaço vetorial, o que ajuda os modelos de IA a inferirem significados contextuais de maneira mais eficaz. Por exemplo, em um espaço de *embeddings*, a distância entre os vetores de "rei" e "rainha" é semelhante à distância entre "homem" e "mulher", capturando relações análogas de gênero e realeza. Essa relação é ilustrada na Figura 2, que apresenta uma visualização de embeddings, mostrando a proximidade e as relações entre as palavras.

2.4.1.2 Desafios e Limitações dos Word Embeddings

Apesar de suas vantagens, os *word embeddings* também têm suas limitações. Um dos principais desafios é que, em modelos mais simples como Word2Vec e GloVe, a representação de uma palavra é fixa, independentemente do contexto em que ela aparece. Isso significa que palavras polissêmicas, que têm múltiplos significados, podem ser mal representadas em algumas frases.

Em um espaço de *embeddings*, palavras como "banco" podem ter múltiplos significados

Figura 3 – Representação estática de palavras polissêmicas



Fonte: Elaborado pelo autor, 2024.

(banco de dados, banco de sentar, instituição financeira). A representação vetorial única não captura essa complexidade, tendendo a representar um significado médio. O significado de uma palavra pode variar dependendo do contexto. Por exemplo, na Figura 3, a seta verde indica uma associação forte entre "banco" e "dinheiro", mas em outros contextos, "banco" poderia estar mais próximo de "cadeira" ou "sentar".

Outro desafio é a escalabilidade dos modelos quando aplicados a grandes quantidades de dados. Embora os *embeddings* reduzam a dimensionalidade, o treinamento massivo ainda pode ser computacionalmente caro. Além disso, palavras raras ou de fora do vocabulário podem não ser bem representadas em modelos tradicionais de *embeddings*.

Os modelos de *embeddings* contextuais de Liu, Kusner e Blunsom (2020), resolvem parte desses problemas, mas à custa de maior complexidade computacional, maiores requisitos de dados para treinamento e ajuste fino.

2.5 TRABALHOS CORRELATOS

O uso de técnicas de extração, transformação e análise de dados aplicadas a documentos de patentes tem atraído crescente atenção na literatura científica, dado seu potencial para subsidiar a inovação tecnológica, apoiar decisões estratégicas e viabilizar soluções baseadas em dados em contextos empresariais e acadêmicos. Essa seção discute trabalhos relevantes que se aproximam da proposta deste projeto, com foco em modelagem de dados, pré-processamento textual, uso de IA e integração com *Data Warehouses*.

2.5.1 Modelagem e Infraestrutura de Dados

O trabalho de Zhang (2020) apresenta uma solução de extração e carga de dados com Python para um *Data Warehouse* modelado em estrela no PostgreSQL. A proposta é similar à deste trabalho em termos de estruturação multidimensional dos dados para facilitar consultas analíticas. Complementarmente, Almeida (2011) avalia o uso de ferramentas open source para construção de *Data Warehouses* de baixo custo, ressaltando a viabilidade técnica de soluções acessíveis, especialmente em ambientes acadêmicos.

Silva (2023) discute o desempenho de bancos relacionais versus orientados a grafos para consultas em bases de citações de patentes. Apesar do potencial do modelo orientado a grafos, o PostgreSQL demonstrou melhor desempenho em operações analíticas, o que reforça sua adoção neste projeto para análises de frequência, associação semântica e evolução temporal.

2.5.2 ETL Distribuído e Automação

O DOD-ETL, proposto por Machado et al. (2019), é uma arquitetura de carga distribuída sob demanda para sistemas de BI quase em tempo real. Apesar de não ser o foco central deste projeto, que trabalha com carga em lote, os princípios de automação, modularidade e isolamento aplicados no DOD-ETL também estão presentes na arquitetura aqui desenvolvida, especialmente pelo uso de Docker e scripts automatizados para orquestrar a extração e carga de dados.

2.5.3 Análise Semântica e Textual de Patentes

A extração de conhecimento a partir de conteúdo textual técnico é um dos principais desafios da análise de patentes. Lee, Kang e Kang (2022) realiza um levantamento abrangente das aplicações de *deep learning* para tarefas como sumarização, classificação automática, detecção de tópicos e extração de palavras-chave. Esses métodos dependem fortemente da qualidade do pré-processamento textual, etapa que envolve desde a remoção de *stopwords* até a normalização e lematização, práticas adotadas também neste trabalho.

Um destaque recente é o sistema EvoPat, apresentado em Wang et al. (2024), que utiliza múltiplos agentes baseados em LLMs para analisar patentes sob diferentes perspectivas: pontos de inovação, métodos de implementação, detalhes técnicos, comparação horizontal com outras patentes e orientação acadêmica. A arquitetura de EvoPat combina pré-processamento textual, análise semântica e geração estruturada de relatórios, armazenando os embeddings das patentes em bases vetoriais. A proposta é especialmente relevante por demonstrar como a combinação de IA e estruturação de dados pode amplificar a compreensão e o valor extraído de documentos complexos como as patentes.

2.5.4 Técnicas Avançadas e Tarefas Específicas

Em relação a tarefas específicas aplicadas à análise de patentes, diversos estudos apontam o papel de técnicas de IA em atividades como:

Sumarização automática: técnicas baseadas em redes neurais têm sido aplicadas para gerar resumos de patentes, tornando seu conteúdo mais acessível e sintetizado para usuários finais (Mihalcea; Tarau, 2004).

Classificação temática: algoritmos supervisionados e métodos de *fine-tuning* em modelos como BERT têm sido utilizados para categorizar automaticamente documentos em áreas técnicas específicas (Fall et al., 2003).

Extração de palavras-chave: métodos como TF-IDF, RAKE e redes semânticas têm sido combinados com embeddings para identificar os termos centrais de uma invenção (Turney, 2000).

Deteção de tendências: a análise temporal da coocorrência de termos técnicos tem sido empregada para mapear o surgimento de tecnologias emergentes e orientar decisões de investimento e pesquisa (Yoon; Park; Kim, 2004).

2.5.5 Síntese e Contribuição do Presente Trabalho

A partir da revisão dos trabalhos correlatos, observa-se uma convergência entre a necessidade de organização estruturada dos dados, vvia modelagem dimensional, a aplicação de técnicas de PLN e IA para enriquecimento semântico, e o uso de pipelines automatizados para lidar com grandes volumes de informações técnicas.

Este trabalho propõe uma abordagem integrada que contempla todas essas frentes: desde a ingestão de arquivos XML da USPTO até a estruturação em um *Data Warehouse* acessível por meio de API REST, com suporte a consultas OLAP e à aplicação futura de modelos inteligentes. Ao fazer isso, contribui para democratizar o acesso e o uso estratégico de dados patentários, conectando engenharia de dados, ciência da informação e inteligência computacional.

3 DESENVOLVIMENTO

Este capítulo detalha o processo de desenvolvimento da arquitetura de *Data Warehouse* automatizado voltada à análise de dados de patentes. O objetivo é transformar grandes volumes de texto técnico em informações estruturadas e úteis para apoiar a gestão do conhecimento e a tomada de decisões estratégicas. Para isso, são descritas as etapas fundamentais do projeto, desde a modelagem do banco de dados até a implementação do *pipeline* de ETL, incluindo o uso de técnicas de pré-processamento textual, extração de entidades técnicas com modelos de linguagem e integração com um modelo dimensional otimizado para consultas analíticas. As decisões de arquitetura e tecnologia são apresentadas com foco na reprodutibilidade, escalabilidade e aderência ao domínio das patentes.

3.1 VISÃO GERAL DA ARQUITETURA

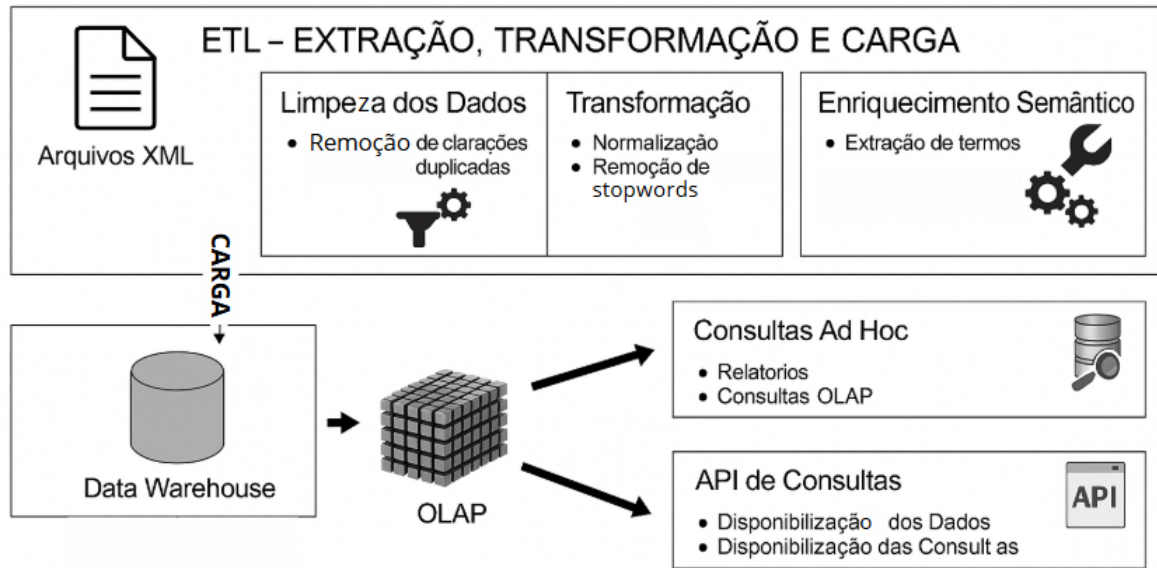
A arquitetura desenvolvida neste trabalho tem como propósito viabilizar, de forma automatizada e escalável, a coleta, transformação, estruturação e análise de dados técnicos provenientes da base de patentes da USPTO. Voltada ao apoio à gestão do conhecimento e à geração de *insights* sobre inovação tecnológica, a solução combina boas práticas de engenharia de dados com recursos de processamento textual e IA, formando uma cadeia integrada de valor informacional.

A proposta é composta por três camadas tecnológicas principais, organizadas de forma modular:

- Um *pipeline* ETL automatizado desenvolvido em Python, responsável por processar os dados desde a extração até sua inserção no modelo analítico;
- Um banco de dados PostgreSQL estruturado como um *Data Warehouse* com modelagem dimensional otimizada para consultas analíticas;
- Um ambiente containerizado com Docker, que assegura reprodutibilidade, isolamento de dependências e facilidade de implantação em diferentes contextos computacionais.

O fluxo de dados tem início com a leitura de arquivos XML contendo registros de pedidos de patentes, os quais são submetidos a diversas etapas de pré-processamento, transformação e enriquecimento semântico incluindo tokenização, filtragem de *stopwords* e extração de entidades técnicas com apoio de modelos de linguagem. Os dados estruturados são então carregados em uma camada intermediária de *staging area* para controle de qualidade e, em seguida, transferidos para as tabelas de dimensão e fato do *Data Warehouse*. A Figura 4 ilustra esse fluxo completo.

Figura 4 – Representação da arquitetura



Fonte: Elaborado pelo autor, 2025.

3.2 MODELAGEM DO DATA WAREHOUSE

A modelagem adotada neste trabalho segue o paradigma dimensional do tipo estrela, conhecido por sua eficiência em consultas analíticas e pela clareza na representação de dados históricos e temáticos. No centro dessa modelagem está a tabela fato *fact_patents*, cuja principal função é registrar a ocorrência de palavras extraídas dos resumos das patentes, associando-as a suas respectivas dimensões contextuais. Cada registro nesta tabela representa a frequência com que uma determinada palavra aparece no resumo de uma patente, permitindo análises semânticas e temporais robustas.

As tabelas de dimensão conectadas a *fact_patents* são responsáveis por armazenar os atributos descritivos necessários para análise. A *dim_authors* armazena os nomes dos inventores; a *dim_countries*, os países de origem dos pedidos de patente; a *dim_date*, os dados temporais estruturados em dia, mês e ano; a *dim_categories*, as classificações técnicas das patentes conforme a IPC¹; a *dim_patents*, os metadados da invenção, incluindo título, número do documento, resumo e descrição; e a *dim_words*, que concentra as palavras únicas extraídas dos resumos, incluindo os termos gerados por modelos de linguagem com o campo *is_generated_term* = True.

Complementando esse modelo dimensional, foi implementada uma área de preparação chamada *staging_patents*, que atua como uma camada intermediária no processo de carga. Nessa etapa, os dados são armazenados de forma bruta ou parcialmente transformada, permitindo validações e enriquecimentos antes de sua inserção definitiva nas tabelas do *Data Warehouse*. Essa estrutura garante integridade, rastreabilidade e flexibilidade na evolução do modelo.

¹ <http://ipc.inpi.gov.br/>

Figura 5 – Modelo dimensional adotado no *Data Warehouse* de patentes.



Fonte: Elaborado pelo autor, 2025.

Esse arranjo não apenas sustenta análises do tipo OLAP, como também possibilita a expansão do sistema com técnicas modernas de Processamento de Linguagem Natural (PLN), como a extração de entidades técnicas com LLMs. A Figura 5 ilustra visualmente essa modelagem, destacando as relações entre a tabela fato e suas respectivas dimensões.

3.3 DESENVOLVIMENTO DO *PIPELINE* ETL

O *pipeline* foi desenvolvido em Python e organizado nos módulos `extract.py`, `transform.py` e `load.py`. Cada fase possui responsabilidade clara e interage com o banco de dados via a biblioteca `psycopg2`.

3.3.1 Extração

A etapa de extração é o ponto de partida do *pipeline* ETL, sendo responsável por coletar os dados brutos a partir dos arquivos XML disponibilizados semanalmente pela base USPTO. Esses arquivos contêm registros completos de patentes publicadas, codificados em uma estrutura hierárquica e padronizada. Esses arquivos, no entanto, contêm múltiplas declarações `<?xml ... ?>`, o que viola a estrutura do XML padrão e impede o *parse* com bibliotecas como `ElementTree`. Para resolver esse problema, desenvolveu-se o módulo `fix_xml.py`, que remove todas as declarações inválidas e encapsula o conteúdo em uma única raiz `<root>`:

Listing 3.1 – Estrutura original de um arquivo baixado da USPTO (inválido)

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE us-patent-application SYSTEM "us-patent-application.dtd" [ ]>
  <us-patent-application>
    patente 1...
  </us-patent-application>
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE us-patent-application SYSTEM "us-patent-application.dtd" [ ]>
  <us-patent-application>
    patente 2...
  </us-patent-application>
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE us-patent-application SYSTEM "us-patent-application.dtd" [ ]>
  <us-patent-application>
    patente 3...
  </us-patent-application>
```

Listing 3.2 – XML após tratamento no `fix_xml.py`

```
<root>
  <us-patent-application>
```

```

    patente 1...
  </us-patent-application>
  <us-patent-application>
    patente 2...
  </us-patent-application>
  <us-patent-application>
    patente 3...
  </us-patent-application>
</root>

```

Após isso, os arquivos são lidos e processados por um script em Python, que percorre os elementos XML e extrai informações relevantes para a análise. Entre os campos extraídos, destacam-se: número do documento, título da invenção, país de origem, data de depósito, nome do inventor, resumo e descrição. Informações classificatórias como seção, símbolo de classe e subclasse também são coletadas e associadas a cada patente, com base na estrutura da classificação internacional de patentes.

Cada arquivo disponibilizado semanalmente pela USPTO possui milhares de registros. Para garantir a rastreabilidade dos dados, o nome do arquivo de origem é coletado, permitindo identificar a origem de cada entrada mesmo após sua carga no sistema. Essa estratégia facilita o controle de duplicação e auditoria dos dados extraídos.

A extração é realizada de forma sequencial, com cada patente sendo transformada em um dicionário estruturado contendo os campos relevantes. Ao final da etapa, os dados estruturados são encaminhados para a transformação, que prepara os registros para a inserção segura e padronizada no *Data Warehouse*.

Abaixo é possível verificar a estrutura xml de uma patente já ajustada para extração:

Listing 3.3 – Trecho do XML de patente após pré-processamento

```

<root>
  <us-patent-application lang="EN" country="US" date-publ="20250522">
    <us-bibliographic-data-application>
      <publication-reference>
        <document-id>
          <country>US</country>
          <doc-number>20250160233</doc-number>
          <kind>A1</kind>
          <date>20250522</date>
        </document-id>
      </publication-reference>
      <application-reference>
        <document-id>
          <country>US</country>
          <doc-number>19035437</doc-number>

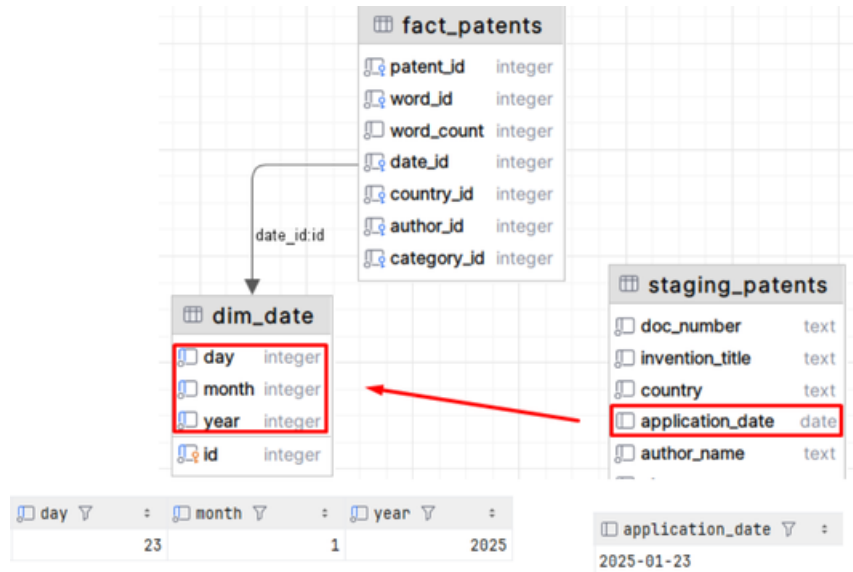
```

```

    <date>20250123</date>
  </document-id>
</application-reference>
<classifications-ipcr>
  <classification-ipcr>
    <section>A</section>
    <class>01</class>
    <subclass>B</subclass>
  </classification-ipcr>
</classifications-ipcr>
<invention-title>APPARATUS FOR COMBINING PLANTING
  IMPLEMENTS</invention-title>
<us-parties>
  <inventors>
    <inventor>
      <addressbook>
        <first-name>Jeffrey Howard</first-name>
        <last-name>Cromwell</last-name>
        <address>
          <city>Camden</city>
          <state>AL</state>
          <country>US</country>
        </address>
      </addressbook>
    </inventor>
  </inventors>
</us-parties>
</us-bibliographic-data-application>
<abstract>
  <p>An apparatus for combining planting implements enables an operator of a
    tractor to use both a cultivating implement and a broadcasting
    implement simultaneously...</p>
</abstract>
<description>
  <p>Hunters often plant food plots to provide feed and nutrition to game
    such as deer and wild turkey. The planting process typically involves
    a cultivating implement and a broadcasting implement...</p>
</description>
</us-patent-application>
</root>

```

Figura 6 – Transformação da data.



Fonte: Elaborado pelo autor, 2025.

3.3.2 Transformação

A etapa de transformação é o núcleo lógico do *pipeline* ETL. Nela, os dados extraídos dos arquivos XML da base USPTO passam por uma série de procedimentos que os tornam aptos para análise no modelo dimensional. O objetivo principal é estruturar, limpar e enriquecer os dados brutos de forma automatizada.

Inicialmente, realiza-se a normalização das datas de depósito. Os valores originais no formato YYYYMMDD são convertidos em objetos de data estruturados (*day*, *month*, *year*), que são posteriormente vinculados à dimensão temporal *dim_date*. Esse relacionamento é fundamental para permitir análises temporais, como a contagem de patentes por ano ou a evolução de termos técnicos ao longo do tempo. A Figura 6 exemplifica essa transformação, onde a data "20250123" é decomposta em 23, 01, 2025 e associada a um registro na *dim_date*.

Em seguida, o nome dos inventores é padronizado por meio da remoção de caracteres especiais e capitalização uniforme, assegurando consistência na dimensão *dim_authors*. Esse processo evita duplicidades causadas por variações na grafia do mesmo nome.

O campo *abstract_text*, que contém o resumo da invenção, é submetido a técnicas de pré-processamento textual. A tokenização é realizada com o auxílio de expressões regulares para extrair apenas palavras válidas do texto. Após essa etapa, todo o texto é convertido para letras minúsculas e é aplicado um filtro de *stopwords* com base na biblioteca `nltk.corpus.stopwords`, removendo palavras comuns da língua inglesa que não agregam valor semântico à análise.

A staging area, representada pela tabela *staging_patents*, exerce um papel estratégico nesse processo. Todos os dados transformados são inicialmente armazenados nesta tabela, permitindo um ponto de controle intermediário entre a extração e a carga definitiva. Essa abordagem facilita a validação de duplicações, a aplicação de enriquecimentos adicionais e a

separação clara entre dados brutos e dados prontos para análise.

Um dos principais diferenciais desta arquitetura é a introdução de enriquecimento semântico por meio de modelos de linguagem de última geração (LLMs). Utilizando o GPT-4o, os campos `invention_title` e `abstract_text` são processados com *prompts* específicos, gerando uma lista de entidades técnicas relevantes para cada patente. Esses termos extraídos são armazenados no campo `generated_terms` da `staging_patents` e posteriormente integrados ao modelo dimensional. A normalização desses termos, como a redução a formas canônicas, remoção de duplicatas e unificação semântica, aumenta significativamente o valor analítico dos dados.

Essas transformações garantem que o conteúdo textual das patentes esteja não apenas limpo e estruturado, mas também semanticamente enriquecido, proporcionando uma base sólida para as análises conduzidas nas etapas seguintes.

Exemplo de resumo antes da transformação:

```
"A_method_and_apparatus_for_secure_communication_in_a_future..."
```

Após transformação:

```
["method", "apparatus", "secure", "communication", "future"]
```

Essas palavras são armazenadas em `dim_words` e relacionadas às suas respectivas patentes na `fact_patents`, com contagem de frequência por palavra.

3.3.3 Carga

A etapa de carga é responsável por inserir os dados transformados nas estruturas permanentes do *Data Warehouse*. Esse processo é dividido em duas fases complementares: a carga inicial na área de staging e a carga final nas tabelas dimensionais e de fatos.

Na primeira fase, os dados transformados são inseridos na tabela `staging_patents`, que atua como uma área intermediária de controle. Essa estratégia permite armazenar os dados de forma segura antes da carga definitiva, possibilitando auditoria, reprocessamentos e verificação de duplicidades. A inserção nessa tabela verifica, por exemplo, se o número do documento já foi previamente carregado a partir do mesmo arquivo fonte, garantindo que não ocorram registros duplicados durante execuções repetidas do pipeline.

A segunda fase consiste na carga incremental para as tabelas dimensionais (`dim_authors`, `dim_countries`, `dim_date`, `dim_categories`, `dim_words`, `dim_patents`) e para a tabela fato `fact_patents`. Essa carga é realizada com uso de chaves substitutas, o que assegura a integridade dos relacionamentos e permite análises do tipo OLAP de forma eficiente.

Cada dimensão é verificada dinamicamente: se um valor ainda não existir, ele é inserido e seu respectivo ID é recuperado para uso na `fact_patents`. Por exemplo, se um novo autor for encontrado no campo `author_name`, seu nome é padronizado e adicionado na `dim_authors`, sendo em seguida vinculado à patente na carga da tabela fato.

O relacionamento com a dimensão temporal também é fundamental. A data de depósito (`application_date`), já normalizada, é decomposta em dia, mês e ano, permitindo a geração de um identificador exclusivo na `dim_date`. Esse identificador é então utilizado para associar temporalmente cada fato registrado no modelo.

A carga na tabela `fact_patents` registra, para cada patente, a ocorrência dos termos extraídos do resumo (palavras simples e entidades compostas), vinculando-os às respectivas dimensões por meio de identificadores substitutos. Essa abordagem garante rastreabilidade, consistência referencial e flexibilidade para consultas analíticas.

O controle do tempo de execução também é monitorado nessa etapa, especialmente no ambiente remoto (Supabase), onde o tempo médio de carga pode variar de acordo com o volume de registros processados. A separação clara entre carga temporária (staging) e carga final (modelo dimensional) contribui para a robustez e a manutenibilidade da arquitetura.

Esse processo garante consistência entre as tabelas e preserva a rastreabilidade dos dados originais.

3.3.4 Execução Local e em Nuvem via Supabase

O sistema foi projetado para funcionar de forma transparente tanto em ambiente local quanto remoto. Utilizando variáveis de ambiente e um módulo centralizado de conexão com o banco de dados, a mesma base de código permite:

- Executar o *pipeline* ETL localmente, com PostgreSQL em container;
- Executar o mesmo pipeline, sem alterações no código, em um banco remoto hospedado no Supabase.

Essa flexibilidade foi alcançada com o uso de arquivos `.env` e configuração dinâmica do módulo de conexão com o banco de dados, permitindo alternar entre diferentes ambientes sem modificações manuais. Além disso, o Supabase provê uma interface web de administração, facilitando testes, inspeção de dados e demonstrações.

Essa abordagem não apenas garante portabilidade e reprodutibilidade, como também aproxima o projeto de um cenário real de implantação em nuvem.

Além do *pipeline* ETL, a API REST também pode ser configurada para se conectar ao Supabase por meio do mesmo arquivo `.env`, bastando alterar os parâmetros de conexão. Isso permite que aplicações externas ou ferramentas de visualização acessem diretamente os dados processados na nuvem, sem a necessidade de manter o banco local ativo.

Dois cenários distintos podem ser facilmente contemplados com essa abordagem:

- Desenvolvimento local: ideal para testes e ajustes rápidos no código, sem depender de conexão com a internet;
- Implantação remota: permite hospedar o banco na nuvem (Supabase) e conectar tanto o ETL quanto a API REST a esse repositório central, viabilizando demonstrações, dashboards e integrações externas em tempo real.

Outro benefício da execução no Supabase é o suporte a mecanismos de autenticação,

controle de permissões e auditoria, além de uma interface web completa para visualização e gerenciamento das tabelas. Isso o torna especialmente útil para projetos acadêmicos, apresentações e colaboração entre equipes.

A execução em nuvem também garante maior persistência dos dados entre execuções, facilitando reprocessamentos e evitando perdas em caso de reinicializações do ambiente local.

3.4 AUTOMAÇÃO COM DOCKER

Para facilitar o empacotamento e a execução do sistema em diferentes ambientes, a arquitetura foi completamente containerizada utilizando Docker e Docker Compose. O ambiente é composto por três serviços principais:

- `postgres_patents`: container do banco de dados PostgreSQL com inicialização automática do schema;
- `etl_patents`: container com o *pipeline* Python que executa automaticamente o processo ETL na inicialização;
- `api_patents`: serviço FastAPI para consulta dos dados já carregados.

Os volumes definidos no *docker-compose* garantem a persistência tanto dos dados quanto do código fonte, assegurando reprodutibilidade. A execução completa do sistema pode ser iniciada com um único comando demonstrado a seguir e também na Figura 7:

```
docker-compose up --build
```

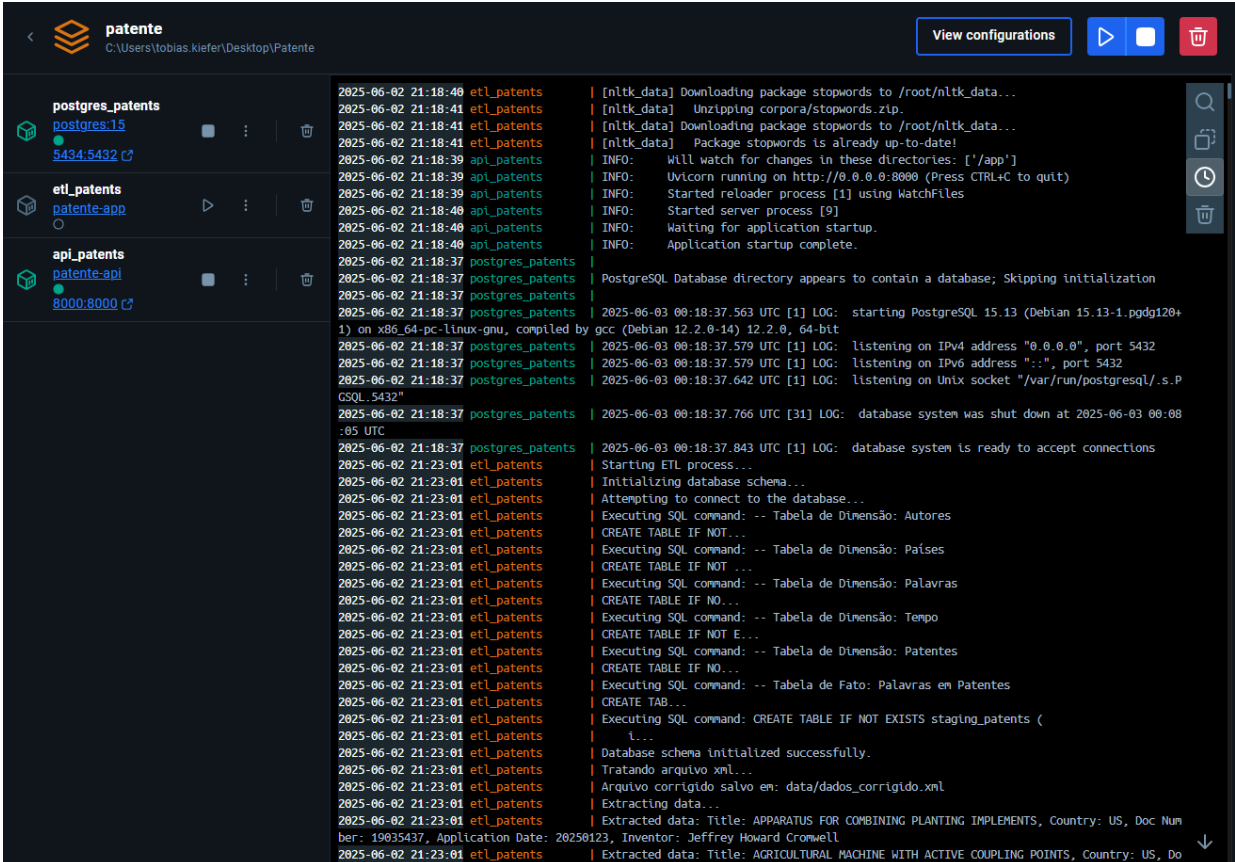
A Figura 7 mostra a execução do ambiente via terminal, enquanto a Figura 8 exibe os logs gerados na interface gráfica do Docker Desktop.

Figura 7 – Execução do ambiente via terminal.

```
(.venv) PS C:\Users\tobias.kiefer\Desktop\Patente> docker-compose up --build
[+] Building 573.6s (16/18)
=> [app internal] load build definition from Dockerfile
=> => transferring dockerfile: 262B
=> [api internal] load build definition from Dockerfile
=> => transferring dockerfile: 262B
=> [api internal] load metadata for docker.io/library/python:3.10-slim
=> [app auth] library/python:pull token for registry-1.docker.io
=> [app internal] load .dockerignore
=> => transferring context: 2B
=> [api internal] load .dockerignore
=> => transferring context: 2B
=> [api 1/4] FROM docker.io/library/python:3.10-slim@sha256:49454d2bf78a48f217eb25ecbcb4b5face313fea6a6e82706465a6990303ada2
=> => resolve docker.io/library/python:3.10-slim@sha256:49454d2bf78a48f217eb25ecbcb4b5face313fea6a6e82706465a6990303ada2
=> [app internal] load build context
=> => transferring context: 2.436B
=> [api internal] load build context
=> => transferring context: 2.436B
=> CACHED [app 2/4] WORKDIR /app
=> [api 3/4] COPY . .
=> [api 4/4] RUN apt-get update && apt-get install -y postgresql-client && pip install --no-cache-dir -r requirements.txt
=> => naming to docker.io/library/patente-app:latest
=> => unpacking to docker.io/library/patente-app:latest
=> [api] exporting to image
=> => exporting layers
=> => exporting manifest sha256:a51476d02973f8aa808375ae99a410396b089645c05b7375f0e83c305387da4c
=> => exporting config sha256:d75891d024935a1cdc0b1421650d3e70ea1890ecee4bd0263c3b0b441ca89938
=> => exporting attestation manifest sha256:ba94a0c1ea148fbc5404b40171fa94bc70e0284e5ff2bb51bd04b86f4bd96c69
=> => exporting manifest list sha256:124e32df34443524c80427c3f0c43b5a01d242ec575f41ef2b8c5a59d14c7dc6
=> => naming to docker.io/library/patente-api:latest
=> => unpacking to docker.io/library/patente-api:latest
=> [app] resolving provenance for metadata file
=> [api] resolving provenance for metadata file
[+] Running 5/5
✓ api Built
✓ app Built
✓ Container postgres_patents Created
✓ Container api_patents Recreated
✓ Container etl_patents Recreated
```

Fonte: Elaborado pelo autor, 2025.

Figura 8 – Logs da execução no Docker Desktop.



The screenshot shows the Docker Desktop interface with the 'patente' container selected. The logs are displayed in a dark-themed window. The container name 'patente' is at the top left, with its path 'C:\Users\tobias.kiefer\Desktop\Patente' below it. A 'View configurations' button is on the top right. The logs are organized into three sections: 'postgres_patents', 'etl_patents', and 'api_patents'. The 'postgres_patents' section shows logs for 'postgres:15' on port 5434:5432. The 'etl_patents' section shows logs for 'patente-app'. The 'api_patents' section shows logs for 'patente-api' on port 8000:8000. The logs themselves are a mix of system messages, INFO messages, and SQL commands. The 'etl_patents' logs show the ETL process starting, including downloading stop words, unzipping corpora, and creating tables. The 'api_patents' logs show the PostgreSQL database starting, including listening on IPv4 and IPv6 addresses and Unix socket.

```

2025-06-02 21:18:40 etl_patents | [nlk_data] Downloading package stopwords to /root/nltk_data...
2025-06-02 21:18:41 etl_patents | [nlk_data] Unzipping corpora/stopwords.zip.
2025-06-02 21:18:41 etl_patents | [nlk_data] Downloading package stopwords to /root/nltk_data...
2025-06-02 21:18:41 etl_patents | [nlk_data] Package stopwords is already up-to-date!
2025-06-02 21:18:39 api_patents | INFO: Will watch for changes in these directories: ['/app']
2025-06-02 21:18:39 api_patents | INFO: Uvicorn running on http://0.0.0.0:8000 (Press CTRL+C to quit)
2025-06-02 21:18:39 api_patents | INFO: Started reloader process [1] using WatchFiles
2025-06-02 21:18:40 api_patents | INFO: Started server process [9]
2025-06-02 21:18:40 api_patents | INFO: Waiting for application startup.
2025-06-02 21:18:40 api_patents | INFO: Application startup complete.
2025-06-02 21:18:37 postgres_patents | PostgreSQL Database directory appears to contain a database; Skipping initialization
2025-06-02 21:18:37 postgres_patents |
2025-06-02 21:18:37 postgres_patents | 1) on x86_64-pc-linux-gnu, compiled by gcc (Debian 12.2.0-14) 12.2.0, 64-bit
2025-06-02 21:18:37 postgres_patents | 2025-06-03 00:18:37.563 UTC [1] LOG: starting PostgreSQL 15.13 (Debian 15.13-1.pgdg120+
2025-06-02 21:18:37 postgres_patents | 2025-06-03 00:18:37.579 UTC [1] LOG: listening on IPv4 address "0.0.0.0", port 5432
2025-06-02 21:18:37 postgres_patents | 2025-06-03 00:18:37.579 UTC [1] LOG: listening on IPv6 address ":::", port 5432
2025-06-02 21:18:37 postgres_patents | 2025-06-03 00:18:37.642 UTC [1] LOG: listening on Unix socket "/var/run/postgresql/.s.P
2025-06-02 21:18:37 postgres_patents | 2025-06-03 00:18:37.766 UTC [31] LOG: database system was shut down at 2025-06-03 00:08
2025-06-02 21:18:37 postgres_patents | .05 UTC
2025-06-02 21:18:37 postgres_patents | 2025-06-03 00:18:37.843 UTC [1] LOG: database system is ready to accept connections
2025-06-02 21:23:01 etl_patents | Starting ETL process...
2025-06-02 21:23:01 etl_patents | Initializing database schema...
2025-06-02 21:23:01 etl_patents | Attempting to connect to the database...
2025-06-02 21:23:01 etl_patents | Executing SQL command: -- Tabela de Dimensão: Autores
2025-06-02 21:23:01 etl_patents | CREATE TABLE IF NOT...
2025-06-02 21:23:01 etl_patents | Executing SQL command: -- Tabela de Dimensão: Países
2025-06-02 21:23:01 etl_patents | CREATE TABLE IF NOT ...
2025-06-02 21:23:01 etl_patents | Executing SQL command: -- Tabela de Dimensão: Palavras
2025-06-02 21:23:01 etl_patents | CREATE TABLE IF NO...
2025-06-02 21:23:01 etl_patents | Executing SQL command: -- Tabela de Dimensão: Tempo
2025-06-02 21:23:01 etl_patents | CREATE TABLE IF NOT E...
2025-06-02 21:23:01 etl_patents | Executing SQL command: -- Tabela de Dimensão: Patentes
2025-06-02 21:23:01 etl_patents | CREATE TABLE IF NO...
2025-06-02 21:23:01 etl_patents | Executing SQL command: -- Tabela de Fato: Palavras em Patentes
2025-06-02 21:23:01 etl_patents | CREATE TAB...
2025-06-02 21:23:01 etl_patents | Executing SQL command: CREATE TABLE IF NOT EXISTS staging_patents (
2025-06-02 21:23:01 etl_patents | t....
2025-06-02 21:23:01 etl_patents | Database schema initialized successfully.
2025-06-02 21:23:01 etl_patents | Tratando arquivo xml...
2025-06-02 21:23:01 etl_patents | Arquivo corrigido salvo em: data/dados_corrigido.xml
2025-06-02 21:23:01 etl_patents | Extracting data...
2025-06-02 21:23:01 etl_patents | Extracted data: Title: APPARATUS FOR COMBINING PLANTING IMPLEMENTS, Country: US, Doc Num
2025-06-02 21:23:01 etl_patents | ber: 19035437, Application Date: 202508123, Inventor: Jeffrey Howard Cromwell
2025-06-02 21:23:01 etl_patents | Extracted data: Title: AGRICULTURAL MACHINE WITH ACTIVE COUPLING POINTS, Country: US, Do

```

Fonte: Elaborado pelo autor, 2025.

3.5 API REST PARA CONSULTA DE DADOS

Como parte opcional e complementar da arquitetura desenvolvida, foi implementada uma API REST para expor os dados analíticos armazenados no *Data Warehouse*, especialmente as consultas OLAP mencionadas anteriormente. O objetivo é permitir o consumo externo das informações processadas pelo *pipeline* ETL, viabilizando visualizações, integrações e análises automatizadas.

A API foi construída utilizando o *framework* FastAPI, escolhido por sua leveza, alto desempenho e pela geração automática de documentação interativa via Swagger UI.

A Figura 16 ilustra o consumo da API por meio da ferramenta Postman, utilizando o endpoint `/words/associadas` com o termo `electric`, retornando palavras semanticamente associadas com base na coocorrência nos resumos das patentes.

3.5.1 Estrutura da API

A API está containerizada em um serviço independente dentro do ambiente Docker Compose, escutando na porta 8000 por padrão.

Sua estrutura organizacional está dividida da seguinte forma:

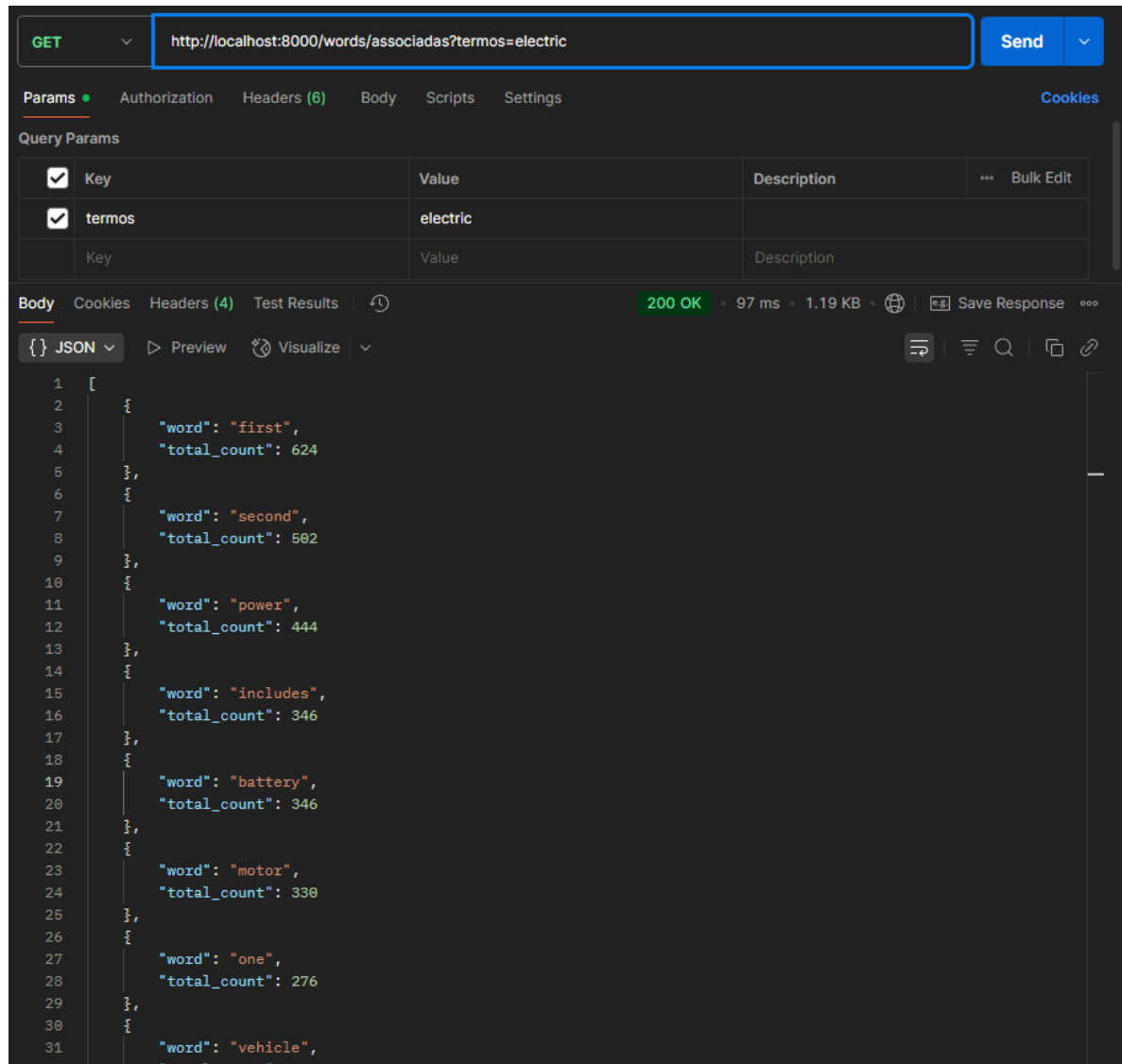
- `main.py`: ponto de entrada da aplicação FastAPI;
- `endpoints/`: diretório com os módulos de rotas agrupadas por domínio:
 - `words.py`: análise de palavras;
 - `authors.py`: informações sobre inventores;
 - `countries.py`: dados agrupados por país.
- `queries.py`: repositório de consultas OLAP reutilizáveis;
- `db_connection.py`: módulo de conexão com o banco de dados PostgreSQL.

3.5.2 Principais Endpoints da API REST

Os principais endpoints disponibilizados pela API são:

- `GET /words/top`: Retorna as 30 palavras mais frequentes nos resumos das patentes.
- `GET /words/por-ano`: Retorna a frequência de palavras agrupada por ano.
- `GET /words/por-pais`: Retorna a frequência de palavras por país de origem.
- `GET /words/por-autor`: Retorna as palavras mais frequentes por autor.
- `GET /words/ranking-anual`: Retorna o ranking das cinco palavras mais frequentes em cada ano.
- `GET /words/associadas?termos=termo1, termo2`: Retorna palavras semanticamente associadas aos termos fornecidos, com base na coocorrência nos resumos.
- `GET /words/associadas-tempo?termos=termo1, termo2`: Similar ao anterior, mas com análise temporal ao longo dos anos.
- `GET /authors/{nome}`: Retorna as patentes vinculadas a um autor específico.

Figura 9 – Consumo da API - Palavras associadas semanticamente à palavra *electric*.



Fonte: Elaborado pelo autor, 2025.

- GET `/countries/{nome}`: Retorna as patentes associadas a um país específico.

A documentação da API, gerada automaticamente pela FastAPI, pode ser acessada localmente após a execução do sistema².

3.5.3 Consultas OLAP na API

As rotas da API são alimentadas por consultas analíticas sobre o modelo dimensional do *Data Warehouse*, utilizando funções como SUM, COUNT, GROUP BY, ORDER BY e, em alguns casos, funções de janela como ROW_NUMBER(). Essas consultas permitem análises complexas sobre a base de patentes.

Entre os tipos de análises possíveis, destacam-se:

² <http://localhost:8000/docs>

- Evolução temporal de termos técnicos;
- Palavras mais relevantes por país ou autor;
- Identificação de tendências emergentes por coocorrência de termos.

3.5.4 Parâmetros Dinâmicos

Dois dos endpoints da API são dinâmicos e permitem a personalização das consultas com base nos termos fornecidos:

- `/words/associadas?termos=motor,energy`
- `/words/associadas-tempo?termos=neural,network`

Essas rotas aplicam lógica de coocorrência para identificar palavras frequentemente associadas aos termos informados, revelando tópicos emergentes e padrões ocultos nos resumos das patentes.

3.5.5 Benefícios da API REST

A API REST agregou importantes vantagens à arquitetura como um todo:

- Flexibilidade: Permite consultas sob demanda conforme a necessidade do usuário;
- Integração: Facilita o consumo de dados por sistemas de visualização e análise externa;
- Escalabilidade: Isola o acesso aos dados do processo de extração e carga;
- Acessibilidade: Geração automática de documentação interativa.

3.6 EXECUÇÃO DO PROJETO

Para executar o sistema completo de ETL e análise de patentes, siga as instruções a seguir.

3.6.1 Pré-requisitos

Certifique-se de que as ferramentas Docker e Docker Compose estejam instaladas no ambiente local.

3.6.2 Clonar o Repositório GitHub

Acesse o repositório oficial do projeto³ e clone-o localmente.

3.6.3 Baixar o Arquivo XML da USPTO

Acesse o portal da USPTO⁴ e faça o download de um arquivo XML com pedidos de patente. Renomeie o arquivo para `dados.xml` e mova-o para o diretório `data/` do projeto.

³ <https://github.com/tobiasfkk/etl-patentes>

⁴ <https://data.uspto.gov/bulkdata/datasets/appxml>

3.6.4 Executar o Sistema com Docker

Na raiz do projeto, execute o seguinte comando:

```
docker-compose up --build
```

Esse comando irá:

- Iniciar o PostgreSQL e aplicar o schema definido em `init_schema.sql`;
- Executar o *pipeline* ETL automaticamente;
- Carregar os dados transformados no *Data Warehouse*;
- Subir a API REST para permitir o consumo dos dados.

4 RESULTADOS E DISCUSSÕES

Este capítulo apresenta os principais resultados obtidos a partir da implementação da arquitetura de *Data Warehouse* automatizado voltado à análise de dados de patentes. A discussão está centrada na validação do modelo dimensional proposto, especialmente no que se refere à sua capacidade de suportar consultas OLAP e alinhada aos objetivos de gestão do conhecimento tecnológico.

São descritas e analisadas as principais consultas realizadas sobre a base carregada, abordando aspectos como desempenho de execução, interpretação dos dados retornados e potencial de geração de valor por meio da extração de informações relevantes. Partindo do consumo OLAP também são discutidas outras formas de consumo e exploração da base, como o uso de técnicas de aprendizado de máquina para classificação, regressão e agrupamento, reforçando a adaptabilidade e aplicabilidade prática da solução em contextos reais de inovação e inteligência competitiva.

4.1 EXECUÇÃO DO PIPELINE ETL

O *pipeline* ETL foi executado em ambiente local utilizando Docker, a partir de arquivos XML disponibilizados pela base USPTO.

Cada lote processado foi validado manualmente por meio da inspeção dos dados inseridos nas tabelas de dimensão e fato. O tempo médio de execução do processo ETL por lote variou entre 4 e 6 minutos para executar localmente e em torno de 2h no Supabase, dependendo do número de registros e da complexidade dos textos contidos nas patentes.

Durante a etapa de transformação, foram extraídos atributos como título da invenção, país, nome do autor, categoria, data de depósito, resumo e descrição. Os resumos foram tokenizados, e as palavras foram inseridas no modelo dimensional com seus respectivos contadores de frequência.

4.2 QUALIDADE E CONSISTÊNCIA DOS DADOS CARREGADOS

A modelagem em estrela garantiu a consistência referencial entre os dados, uma vez que todas as inserções foram validadas em tempo de execução por meio de verificação prévia da existência de registros em tabelas de dimensão.

A criação de dimensões normalizadas como *dim_authors*, *dim_countries* e *dim_date* possibilitou a padronização e o cruzamento das informações, facilitando a análise posterior.

O volume total de palavras extraídas e associadas às patentes variou conforme o conteúdo textual de cada documento, e as palavras foram armazenadas com preservação da granularidade textual, mesmo em casos de repetições ou variações morfológicas.

4.3 EXTRAÇÃO DE ENTIDADES TÉCNICAS COM APOIO DE LLMS

Com o objetivo de enriquecer semanticamente os dados textuais armazenados na *staging area*, foi implementado um processo automatizado de extração de termos técnicos utilizando um modelo de linguagem de última geração (LLM), especificamente o GPT-4o da OpenAI. Essa técnica insere uma camada adicional de pré-processamento, na qual o conteúdo dos campos *invention_title* e *abstract_text* é analisado por meio de *prompts* especializados com regras de filtragem, padronização e seleção.

O resultado do processo é gravado no campo *generated_terms* da tabela *staging_patents*, em formato de lista, conforme ilustrado abaixo:

Listing 4.1 – Consulta de patente com termos gerados

```
SELECT
invention_title,
abstract_text,
generated_terms
FROM
staging_patents
WHERE
generated_terms IS NOT NULL
LIMIT 1;
```

Listing 4.2 – Exemplo de termos técnicos extraídos

```
[
"battery_management_system",
"power_inverter",
"electric_vehicle_motor",
"charging_station_interface",
"regenerative_braking"
]
```

Esse enriquecimento semântico permite análises mais profundas sobre os conceitos descritos nos resumos das patentes, superando as limitações das abordagens baseadas apenas em palavras frequentes.

4.4 VALIDAÇÃO DO MODELO MULTIDIMENSIONAL E CONSULTAS OLAP

A validação do modelo multidimensional desenvolvido neste trabalho foi realizada por meio da execução de um conjunto de consultas analíticas baseadas em OLAP. O objetivo central dessa etapa é demonstrar a capacidade da arquitetura proposta em oferecer suporte à exploração de dados de patentes sob diferentes perspectivas como tempo, autor, país e termos técnicos,

permitindo análises consistentes e alinhadas às necessidades de extração de conhecimento estratégico.

As consultas foram definidas de modo a simular cenários reais de exploração de dados, como identificação de tecnologias emergentes, mapeamento de especializações por país, análise de produtividade por autor e evolução temporal de termos técnicos. Para isso, foram utilizados filtros, agrupamentos e métricas aplicadas diretamente sobre a estrutura estrela do *Data Warehouse*.

4.4.1 Integração com o Modelo Dimensional

Após a geração automatizada dos termos técnicos, os dados foram integrados ao modelo dimensional do *Data Warehouse*. Cada termo extraído foi verificado na dimensão `dim_words`, sendo inserido com o atributo `is_generated_term = TRUE` quando ainda inexistente. Em seguida, foi realizada a contagem de ocorrências de cada termo no conteúdo do campo `abstract_text`, considerando o número de vezes que o termo aparece de forma exata e normalizada.

A partir dessas informações, foram identificados os IDs das dimensões relacionadas como país, autor, data e categoria da patente e os termos foram associados às respectivas patentes na tabela `fact_patents`. Esse procedimento garante que as entidades técnicas geradas estejam representadas de forma consistente e integrada no modelo, permitindo consultas cruzadas e análises temporais sobre o uso e a recorrência de conceitos técnicos.

Essa estratégia reflete a mesma abordagem adotada para palavras simples extraídas do resumo, assegurando uniformidade na estrutura e potencializando as análises OLAP. Com isso, torna-se possível comparar, filtrar ou correlacionar palavras frequentes com termos compostos e entidades técnicas relevantes, ampliando a capacidade analítica da base e apoiando decisões baseadas em conhecimento tecnológico estruturado.

Com a base carregada, foram elaboradas consultas OLAP que permitiram análises descritivas e exploratórias. Entre os principais resultados, destacam-se:

- As palavras mais frequentes nos resumos de patentes;
- O ranking anual das palavras mais utilizadas, permitindo a identificação de tendências ao longo do tempo;
- A associação semântica entre termos técnicos recorrentes, por meio de coocorrência em patentes com termos-chave fornecidos;
- A evolução temporal dessas associações, revelando padrões de emergência e declínio de certas expressões ao longo dos anos.

Tais consultas permitiram visualizar a distribuição e a dinâmica das palavras técnicas no corpus de patentes, viabilizando *insights* úteis sobre o foco temático das invenções ao longo do tempo e entre diferentes contextos geográficos ou institucionais.

4.4.2 Parâmetros Experimentais

Os experimentos realizados para validar o modelo dimensional proposto foram conduzidos com base em uma amostra real de dados de 15399 patentes. Vale destacar que os dados analisados neste trabalho foram extraídos a partir de apenas dois arquivos XML disponibilizados pela USPTO, correspondentes às datas de 29/05/2025 e 02/01/2020. Esses arquivos representam semanas específicas de publicações e contemplam documentos de patentes com datas de depósito entre os anos de 2014 e 2025.

A Tabela 1 apresenta a distribuição dessas patentes de acordo com as seções da IPC

Tabela 1 – Seções com Maior Número de Patentes e suas Principais Subclasses

Seção	Total de Patentes	Principais Subclasses
Eletricidade	4452	Circuitos Eletrônicos: 1597 patentes Transmissão de Sinais: 578 patentes Dispositivos de Controle: 505 patentes Medição; Testes (C12Q): 354 patentes Computação e Processamento de Dados: 267 patentes
Física	4401	Medição; Testes: 1515 patentes Instrumentação Científica: 455 patentes Computação e Processamento de Dados: 432 patentes Cuidados Médicos ou Higiene Pessoal: 301 patentes Circuitos Eletrônicos: 251 patentes
Necessidades Humanas	2309	Instrumentação Científica: 639 patentes Cuidados Médicos ou Higiene Pessoal: 589 patentes Medição; Testes: 245 patentes Computação e Processamento de Dados: 218 patentes Medição; Testes (C12Q): 176 patentes

Fonte: Elaborado pelo autor, 2025.

A partir desse agrupamento, foi possível identificar as subclasses mais representativas dentro do conjunto de dados analisado. A Tabela 2 apresenta as 10 subclasses IPC com maior número de patentes, evidenciando áreas tecnológicas de destaque. Essa análise permite observar a concentração de inovações técnicas em determinadas áreas, como Engenharia Mecânica, Tecnologias da Informação e Instrumentação Médica, reforçando o papel dessas categorias na estrutura temática das patentes registradas.

Após a execução do pipeline de ETL, os dados foram estruturados em um *Data Warehouse* com modelo dimensional em estrela, composto por uma tabela fato central (*fact_patents*) e cinco tabelas de dimensão: *dim_words*, *dim_date*, *dim_authors*, *dim_countries* e *dim_patents*. Com base nessa estrutura, foi possível realizar consultas OLAP otimizadas, aplicando filtros por tempo, autor e termo técnico, que serviram de base para as análises exploratórias apresentadas nas subseções seguintes.

Para cada consulta apresentada, foram realizadas dez execuções consecutivas, e os tempos médios ou medianas de resposta foram registrados com o intuito de avaliar o desempenho da

Tabela 2 – Ranking das Subclasses IPC com Maior Número de Patentes

Subclasse	Descrição	Total de Patentes
F	Engenharia Mecânica; Iluminação; Aquecimento	2080
B	Técnicas Industriais Diversas	1948
L	Tecnologias de Informação ou Comunicação	1854
K	Instrumentos Médicos ou Cirúrgicos	1319
N	Computação; Redes de Comunicação	1184
D	Têxteis; Papel	966
C	Química; Metalurgia	818
M	Medição; Testes Elétricos ou Ópticos	721
W	Não definida	647
Q	Não definida	577

Fonte: Elaborado pelo autor, 2025.

solução e a viabilidade de seu uso em ambientes reais de análise de dados tecnológicos.

Esta seção descreve os parâmetros utilizados para a execução das consultas no modelo proposto. Para isso, foi utilizada uma base contendo os dados extraídos dos arquivos XML do USPTO, referentes às semanas de 29/05/2025 e 02/01/2020. Após a execução do pipeline ETL, a base resultante foi armazenada em um *Data Warehouse* estruturado com modelo dimensional em estrela, contendo tabelas de dimensão para autores, países, datas, termos técnicos e patentes.

4.4.3 Experimento 1: Análise Temporal de Termos Técnicos por Ano

A primeira análise exploratória foi conduzida com o objetivo de identificar os termos técnicos mais recorrentes nos resumos das patentes ao longo do tempo. Para isso, foi realizada uma consulta OLAP com base na tabela fato (fact_patents), considerando as dimensões dim_words (termos técnicos) e dim_date (tempo). A agregação foi realizada por ano de depósito, e apenas termos classificados como gerados automaticamente (is_generated_term = TRUE) e com presença em ao menos três anos distintos foram incluídos na análise, a fim de evitar termos esporádicos ou não representativos.

Os cinco termos mais frequentes foram selecionados para compor a Tabela 3 e o gráfico correspondente. Esses resultados oferecem uma visão geral do comportamento dos principais termos técnicos ao longo do período analisado (2015 a 2025), permitindo identificar tendências tecnológicas recorrentes na base.

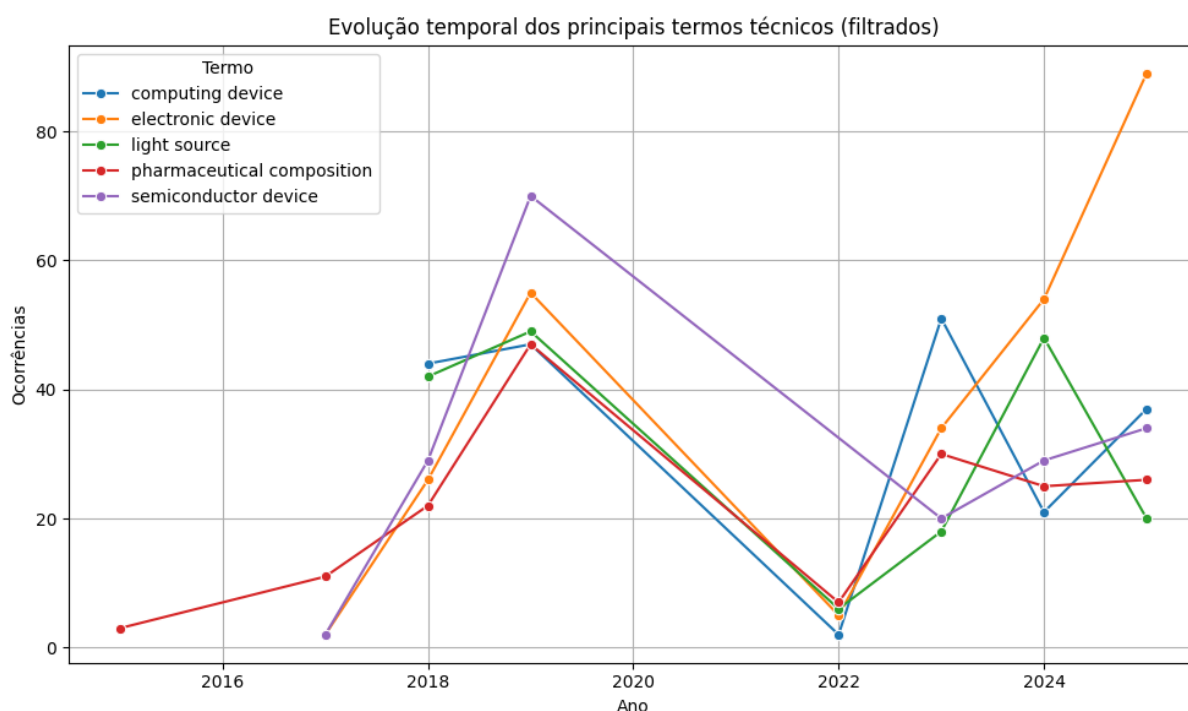
A Tabela 3 apresenta a distribuição temporal das ocorrências dos cinco termos técnicos mais frequentes ao longo dos anos de 2015 a 2025. A consulta foi estruturada sobre a tabela fato fact_patents, utilizando como dimensões principais dim_words (que armazena os termos técnicos extraídos automaticamente dos resumos), e dim_date, para o agrupamento por ano de depósito. O critério de filtragem considerou apenas termos classificados como is_generated_term = TRUE, e com presença em pelo menos três anos distintos, o que assegura a relevância estatística dos termos analisados.

Tabela 3 – Ocorrência temporal dos termos técnicos mais frequentes (2015–2025)

Termo	2015	2017	2018	2019	2020	2022	2023	2024	2025
computing device	0	44	47	22	2	5	51	21	37
electronic device	0	26	55	8	5	6	34	54	89
light source	0	12	43	16	8	3	18	48	20
pharmaceutical composition	3	11	22	4	7	8	30	25	26
semiconductor device	0	23	70	0	20	6	29	20	34

Fonte: Elaborado pelo autor, 2025.

Figura 10 – Evolução temporal dos principais termos técnicos extraídos de patentes (2015–2025)



Fonte: Elaborado pelo autor, 2025.

A análise evidencia um crescimento progressivo no uso do termo *electronic device*, especialmente a partir de 2023, alcançando pico em 2025. O termo *semiconductor device* também apresenta destaque em 2018, ainda que com oscilações nos anos seguintes. Termos como *pharmaceutical composition* e *light source* mostram comportamento mais estável ao longo do tempo. Esses padrões sugerem o surgimento de focos específicos de inovação tecnológica, particularmente em áreas de eletrônica e saúde, refletindo tendências relevantes para estudos de prospecção tecnológica.

A Figura 10 complementa a Tabela 3 ao ilustrar visualmente a evolução dos principais termos técnicos ao longo dos anos. Observa-se um crescimento expressivo no uso do termo *electronic device*, especialmente a partir de 2023, sugerindo um aumento no volume de inovações relacionadas a dispositivos eletrônicos nos documentos analisados. Já o termo *computing device* apresenta oscilações ao longo do período, com pico relevante em 2023. Por sua vez, termos como

pharmaceutical composition e *light source* mantêm uma frequência mais constante, enquanto *semiconductor device* destaca-se pontualmente em 2018. Essas tendências reforçam a importância de análises longitudinais na identificação de domínios tecnológicos emergentes e na compreensão da dinâmica de inovação ao longo do tempo.

Em seguida, análises adicionais são conduzidas com o acréscimo de filtros por outras dimensões como a subclasse tecnológica (*dim_categories*) e país de origem (*dim_countries*) a fim de refinar os padrões observados. Essa abordagem comparativa entre análise geral e análises filtradas possibilita uma compreensão mais profunda dos domínios de inovação e sua distribuição temporal.

4.4.3.1 Explorando Dimensões – Categoria

Além da análise temporal geral, é possível aprofundar a exploração dos dados técnicos ao aplicar filtros por subclasse tecnológica, conforme registrado na dimensão *dim_categories*. A Figura 11 e a Tabela 4 ilustram a evolução dos principais termos técnicos extraídos de patentes associadas à subclasse *Engenharia Mecânica; Iluminação; Aquecimento*, entre os anos de 2017 e 2025.

A análise foi conduzida utilizando a mesma lógica da análise geral: foram selecionados os cinco termos técnicos com maior frequência, desde que classificados como *is_generated_term* = TRUE e com presença em pelo menos três anos distintos. Esses termos foram então agrupados por ano de depósito com base na tabela *fact_patents*, permitindo traçar sua evolução temporal dentro da subclasse selecionada.

A Figura 11 revela padrões distintos dos observados na análise geral. O termo *electronic device* apresenta crescimento contínuo ao longo do período analisado, sugerindo maior relevância no contexto da subclasse avaliada. Já termos como *user interface* e *memory device* mostram picos específicos em determinados anos, indicando o surgimento pontual de tecnologias correlacionadas.

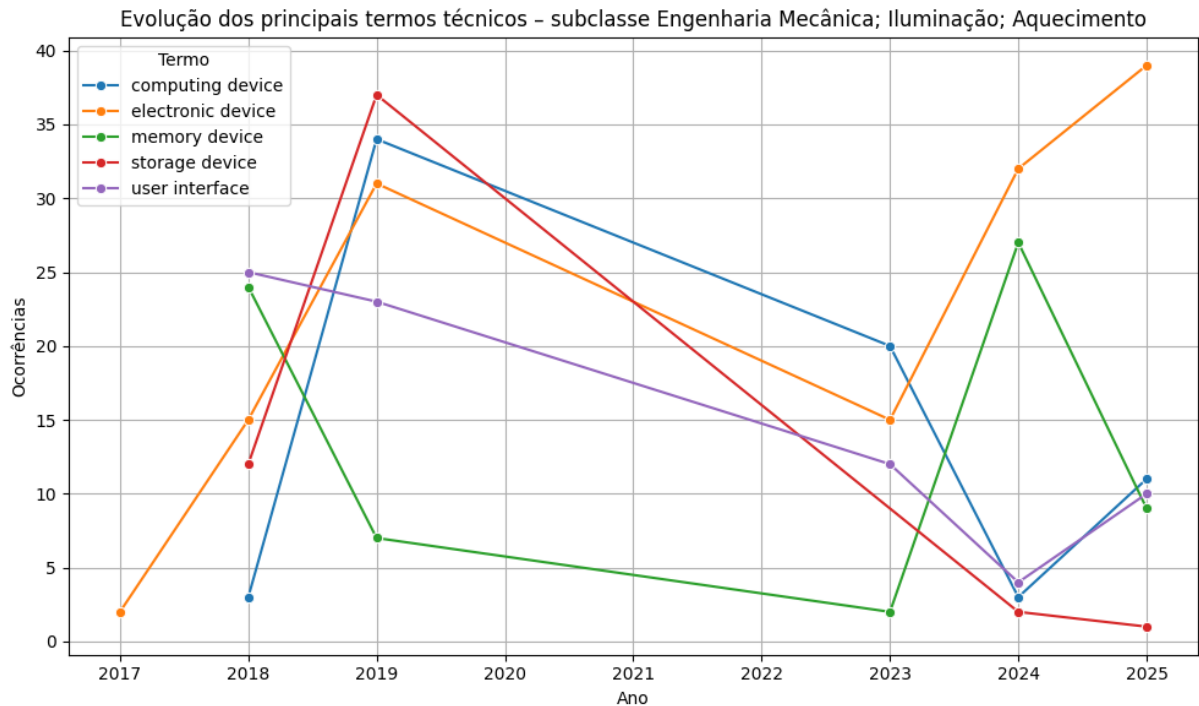
Essa filtragem por categoria tecnológica permite direcionar as análises para áreas temáticas específicas, contribuindo para estudos de inteligência competitiva mais precisos e segmentados por domínio de inovação.

Tabela 4 – Evolução dos principais termos técnicos – subclasse Engenharia Mecânica; Iluminação; Aquecimento

Termo	2017	2018	2019	2023	2024	2025
computing device	0	3	34	20	3	11
electronic device	2	15	31	15	32	39
memory device	0	24	7	2	27	9
storage device	0	12	37	0	2	1
user interface	0	25	23	12	4	10

Fonte: Elaborado pelo autor, 2025.

Figura 11 – Evolução temporal dos principais termos técnicos extraídos de patentes da subclasse Engenharia Mecânica; Iluminação; Aquecimento (2017–2025)



Fonte: Elaborado pelo autor, 2025.

4.4.3.2 Explorando Dimensões – Autores

A análise da distribuição temporal de termos técnicos por autor oferece uma perspectiva relevante sobre o envolvimento de indivíduos ou grupos de inventores em determinados domínios tecnológicos. Esse tipo de investigação permite identificar padrões de especialização, trajetórias de pesquisa e eventuais mudanças no foco de atuação dos autores ao longo do tempo.

Entretanto, devido à limitação do conjunto de dados utilizado neste trabalho, extraído de apenas duas semanas de publicações da base USPTO, não foi possível gerar visualizações significativas para esta dimensão. A maioria dos autores apresenta ocorrências restritas a um ou dois anos, o que inviabiliza a identificação de tendências consistentes ou recorrentes. Mesmo assim, a estrutura do modelo dimensional permanece adequada para esse tipo de análise, devendo se mostrar mais eficaz com o aumento da amostragem de dados em futuros ciclos de execução do pipeline.

4.4.3.3 Explorando Dimensões – Países

A dimensão geográfica, representada na base de dados pela tabela `dim_countries`, também desempenha papel essencial na identificação de tendências tecnológicas e na comparação entre diferentes contextos regionais de inovação. Consultas agrupadas por país podem revelar concentrações temáticas, áreas estratégicas de desenvolvimento e padrões de colaboração internacional.

No entanto, a base utilizada neste experimento, extraída exclusivamente de documentos da USPTO, contempla majoritariamente registros oriundos dos Estados Unidos, o que restringe a diversidade geográfica da análise. Assim, não foi possível realizar comparações significativas entre países nesta etapa. Ainda assim, a arquitetura está preparada para suportar essa análise em cenários futuros, especialmente quando integrada a fontes complementares ou a conjuntos mais amplos de arquivos da própria USPTO.

4.4.4 Associação entre Termos Técnicos

A análise de coocorrência de termos técnicos permite identificar tecnologias frequentemente relacionadas nos mesmos registros de patente. Esse tipo de associação é útil para revelar convergências temáticas, complementaridades entre inovações e áreas potenciais de interdisciplinaridade. Nesta etapa, foram selecionados os cinco termos técnicos mais representativos do modelo dimensional, considerando apenas aqueles com presença em dois ou mais anos distintos. A partir deles, foram identificados os cinco termos mais associados em registros de patente compartilhados.

A Tabela 5 apresenta os resultados dessa análise. Observa-se, por exemplo, que o termo *electronic device* está frequentemente relacionado a conceitos como *storage medium*, *computer program code* e *heart rate information*, sugerindo aplicações em dispositivos inteligentes e monitoramento biométrico. Já *computing device* aparece associado a termos ligados à agricultura de precisão e saúde digital, como *crop yield* e *autism spectrum disorder*.

Essas associações foram extraídas com base na tabela `fact_patents`, relacionando pares de termos técnicos marcados como `is_generated_term = TRUE` e que aparecem em conjunto na mesma patente. Essa abordagem preserva a coerência semântica das associações e respeita o modelo multidimensional ao cruzar informações por patente.

A partir dessa análise, é possível também aplicar filtros adicionais para explorar a associação entre termos dentro de subconjuntos específicos, como categorias técnicas, países de origem ou faixas temporais. Por exemplo, ao restringir as consultas à subclasse *Engenharia Mecânica; Iluminação; Aquecimento*, seria possível observar como os termos associados a *electronic device* variam ao longo dos anos. Essa capacidade de filtragem por múltiplas dimensões reforça o valor analítico do modelo proposto e sua adequação para estudos de prospecção tecnológica em contextos reais.

4.4.5 Associação de Termos Técnicos por Subclasse Tecnológica

Além da análise geral de coocorrência, foi realizada uma consulta com foco na subclasse tecnológica *Engenharia Mecânica; Iluminação; Aquecimento* (código F na classificação IPC). O objetivo foi identificar relações semânticas específicas dentro de um domínio técnico bem definido, reforçando o potencial multidimensional do modelo proposto.

A Tabela 6 apresenta os cinco termos mais fortemente associados aos cinco principais ter-

Tabela 5 – Associação entre termos técnicos com base em coocorrência em patentes

Termo Base	Termo Associado	Ocorrências
<i>electronic device</i>	storage medium	9
	physical distance	4
	computer program code	4
	heart rate information	4
	heart rate measurement sensor	4
<i>computing device</i>	crop yield	4
	computer program product	4
	autism spectrum disorder	4
	crop pour rate	4
	delivered sugar weight	4
<i>semiconductor device</i>	gate structure	11
	dielectric layer	6
	gate electrode	6
	insulating layer	5
	manufacturing method	5
<i>light source</i>	optical device	6
	light sensor	5
	hybrid lighting system	4
	connection port	4
	display device	4
<i>pharmaceutical composition</i>	pharmaceutically acceptable salt	8
	active ingredient	7
	crystalline form	5
	metabolic disorder	5
	pharmaceutically acceptable carrier	5

Fonte: Elaborado pelo autor, 2025.

mos técnicos dessa subclasse, considerando apenas aqueles marcados como `is_generated_term = TRUE`. A associação foi determinada com base na coocorrência dos termos dentro de uma mesma patente, extraída da tabela `fact_patents`.

Os resultados revelam padrões semânticos relevantes. Por exemplo, o termo *storage device* aparece associado a entidades como *machine learning module* e *computer program product*, sugerindo um ecossistema tecnológico voltado ao armazenamento e processamento inteligente de dados. Já *user interface* está fortemente ligado a expressões como *large language model* e *user interaction*, indicando uma interface entre sistemas de IA e seus usuários.

Essa análise destaca como a aplicação de filtros por subclasse tecnológica permite extrair conhecimentos mais granulares sobre as áreas de inovação, reforçando a utilidade da arquitetura proposta como suporte à prospecção tecnológica orientada por contexto.

Tabela 6 – Associação entre termos técnicos – Subclasse Engenharia Mecânica; Iluminação; Aquecimento

Termo Base	Termo Associado	Ocorrências
<i>computing device</i>	storage device	3
	user interface	3
	failure resilient address space	2
	dynamic random access memory	2
	assembly program	1
<i>electronic device</i>	storage medium	4
	flexible display	2
	wearable device	2
	user interface	2
	communication module	2
<i>memory device</i>	bit line	2
	memory module	2
	memory array	2
	memory block	2
	memory controller	2
<i>storage device</i>	computer program product	4
	computing device	3
	machine learning module	3
	output value	2
	error check frequency	2
<i>user interface</i>	computing device	3
	large language model	2
	electronic device	2
	user interaction	2
	computer system	2

Fonte: Elaborado pelo autor, 2025.

4.4.6 Análise Temporal de Tendências Tecnológicas com Modelos Preditivos

Com o objetivo de explorar a capacidade analítica do modelo dimensional proposto, este experimento teve como foco a previsão de termos técnicos recorrentes ao longo do tempo, por meio do consumo de dados estruturados em uma arquitetura de *Data Warehouse*. Inicialmente, foi realizada uma regressão linear simples para estimar a tendência do termo *electronic device*, utilizando como entrada sua frequência anual registrada no modelo dimensional.

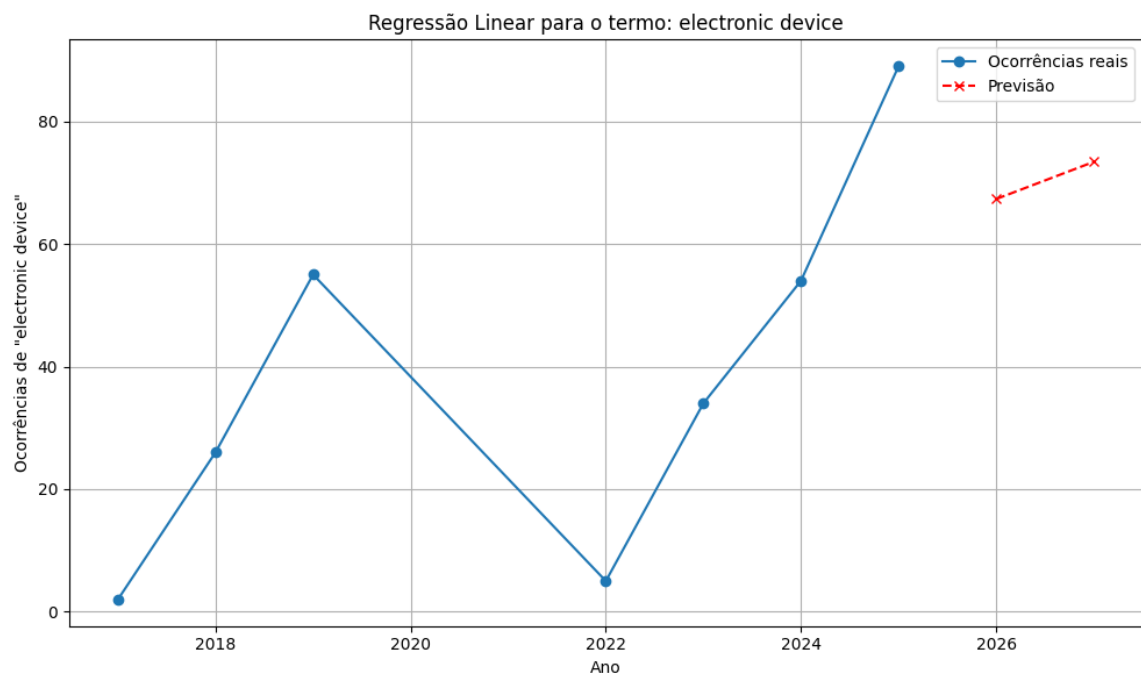
A consulta foi executada sobre a tabela *fact_patents*, relacionando os anos na dimensão *dim_date* com as ocorrências do termo na dimensão *dim_words*. Os dados considerados abrangem o período de 2017 a 2025. A Tabela 7 apresenta as ocorrências reais por ano, bem como as previsões geradas para os anos seguintes.

A Figura 12 ilustra os dados reais e os valores previstos. Nota-se uma tendência crescente a partir de 2023, sugerindo o fortalecimento de tecnologias relacionadas a dispositivos eletrônicos no contexto de inovações registradas em patentes.

Tabela 7 – Ocorrências anuais do termo *electronic device* e previsões futuras

Ano	Ocorrências
2017	2
2018	26
2019	55
2022	5
2023	34
2024	54
2025	89
2026 (previsto)	67
2027 (previsto)	73

Fonte: Elaborado pelo autor, 2025.

Figura 12 – Regressão linear das ocorrências do termo *electronic device* (2017–2027)

Fonte: Elaborado pelo autor, 2025.

Apesar de fornecer uma visão geral da tendência, a regressão linear não considera variações locais nos dados, como picos ou quedas abruptas. Para explorar essas dinâmicas de forma mais sensível, o experimento foi expandido com o uso de uma rede neural artificial do tipo MLP (Multilayer Perceptron), integrada a uma abordagem de janela temporal deslizante.

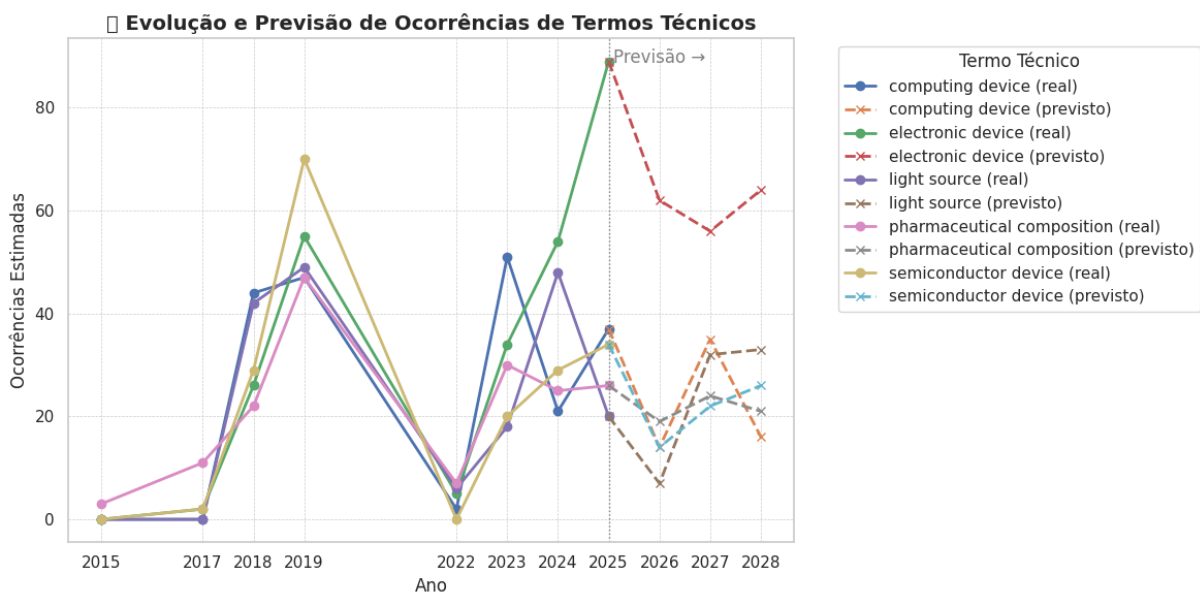
Essa rede neural foi alimentada com séries temporais extraídas diretamente do modelo dimensional, mantendo a estrutura em estrela com a tabela fato `fact_patents` e as dimensões `dim_words`, `dim_date`, `dim_patents`, `dim_authors` e `dim_countries`. Foram selecionadas as 10 palavras técnicas mais frequentes, desde que presentes em pelo menos três anos distintos, garantindo uma base suficientemente densa para aprendizado.

Para cada termo, os dados anuais foram normalizados com `MinMaxScaler`, e a base de treino foi formada por janelas de três anos consecutivos. Por exemplo, para o termo *electronic device*, o vetor de entrada poderia ser $[34, 54, 89]$, representando os anos de 2023 a 2025, cuja previsão seria o valor de 2026.

A arquitetura da rede é composta por uma camada de entrada com três neurônios (referentes à janela temporal), uma camada oculta com cinco neurônios e ativação ReLU, e uma camada de saída com um único neurônio. Após o treinamento, os valores previstos são desnormalizados para comparação com os dados reais.

A Figura 13 mostra o resultado final da aplicação da rede MLP sobre os 10 termos técnicos selecionados. Os valores reais são representados por linhas contínuas, enquanto as previsões para os anos futuros (2026 a 2028) aparecem como linhas tracejadas. Uma linha vertical marca o ponto de transição entre histórico e previsão.

Figura 13 – Evolução e previsão de ocorrência dos principais termos técnicos



A Figura 13 apresenta o comportamento temporal de ocorrência dos dez termos técnicos mais frequentes nas patentes analisadas, incluindo tanto os dados históricos quanto as previsões

geradas pela rede neural.

As linhas contínuas representam os dados reais observados, extraídos diretamente do modelo dimensional, ou seja, o número de vezes que cada termo apareceu em patentes registradas entre os anos analisados. Esses dados históricos formam a base de conhecimento do modelo, permitindo que ele reconheça padrões de crescimento, declínio ou estabilidade ao longo dos anos.

Já as linhas tracejadas indicam os valores previstos para os anos futuros. Essas previsões são o resultado do aprendizado da rede neural sobre as janelas temporais anteriores, ou seja, sequências de três anos consecutivos que alimentam o modelo com contexto suficiente para estimar o comportamento subsequente. Por exemplo, se um termo técnico apresentou aumento consistente de ocorrências entre três anos seguidos, o modelo tende a projetar uma continuação desse crescimento, salvo oscilações aprendidas em padrões semelhantes no histórico.

A separação visual entre os dados reais e os previstos reforça a transição entre aquilo que já é conhecido e aquilo que é estimado com base nas tendências aprendidas. Observa-se que, mesmo com variações individuais, os termos seguem padrões de comportamento compatíveis com sua evolução anterior, demonstrando a capacidade da rede em adaptar-se a diferentes ritmos de crescimento e estabilização.

4.4.7 Agrupamento de Patentes com K-Means e Análise de Clusters Tecnológicos

Neste experimento, aplicou-se a técnica de agrupamento *K-Means* sobre os termos técnicos extraídos das patentes, com o objetivo de identificar subconjuntos de documentos com perfis tecnológicos semelhantes. A consulta foi realizada a partir das tabelas `fact_patents`, `dim_words` e `dim_patents`, utilizando exclusivamente os termos gerados automaticamente por LLMs (`is_generated_term = TRUE`).

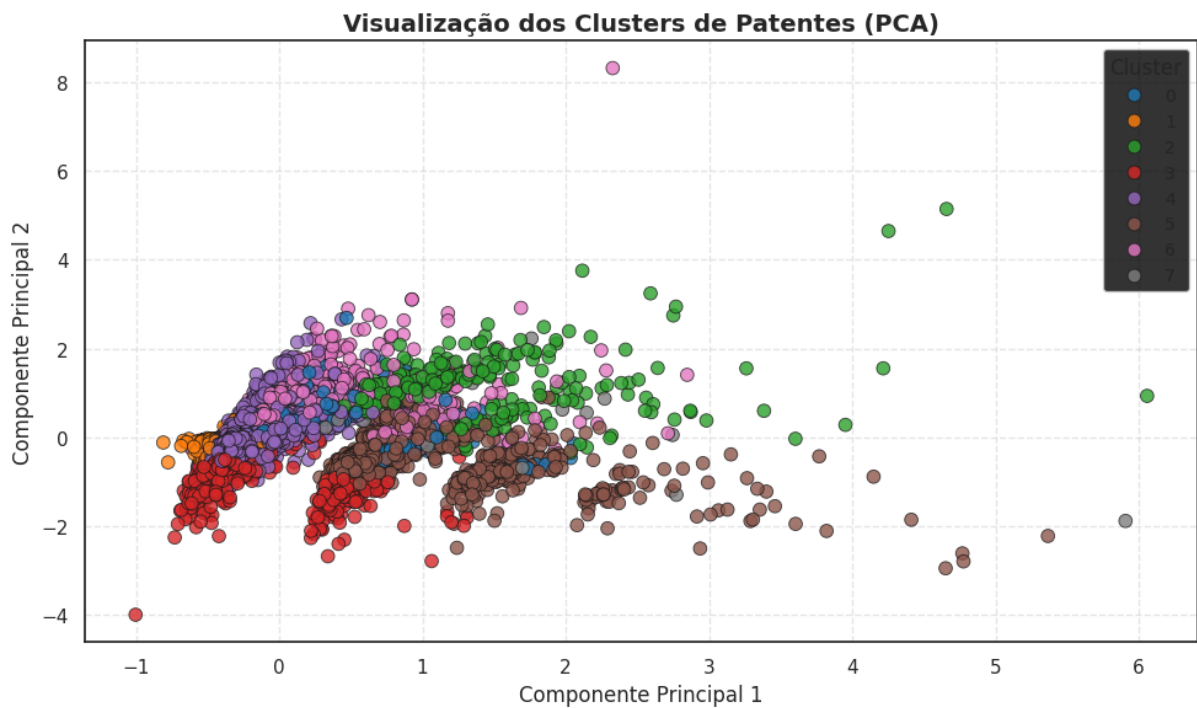
Cada patente foi representada como um vetor de presença de termos — modelo conhecido como *bag-of-words* — e o algoritmo *K-Means* foi executado com a configuração de oito clusters. A seguir, foram aplicadas técnicas de redução de dimensionalidade utilizando *Principal Component Analysis* (PCA), possibilitando a visualização da distribuição espacial dos documentos em duas dimensões.

A Figura 14 apresenta a dispersão das patentes ao longo de dois eixos principais (Componente 1 e Componente 2), evidenciando a formação de agrupamentos bem definidos. Cada ponto representa uma patente, e a cor indica o cluster ao qual ela pertence. Esse tipo de visualização permite observar se há sobreposição entre os grupos e o grau de separação entre as categorias técnico-temáticas.

Clusters mais compactos indicam alta similaridade entre os documentos, enquanto dispersões maiores sugerem diversidade técnica dentro daquele grupo. A separação entre os grupos reforça a eficácia da técnica para a identificação de domínios tecnológicos distintos a partir dos termos extraídos.

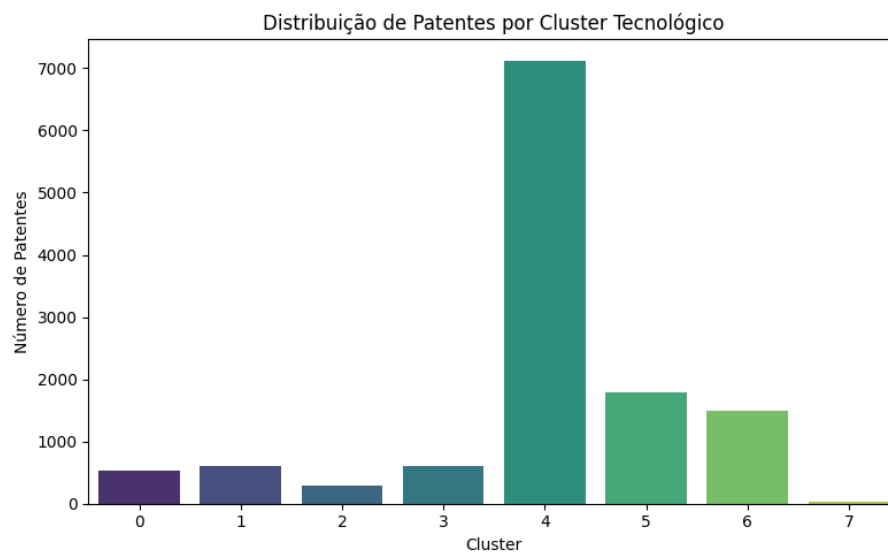
Além da visualização espacial, a contagem de documentos por cluster permite entender a

Figura 14 – Visualização dos clusters de patentes em duas dimensões (redução via PCA).



predominância de determinadas áreas técnicas na amostra analisada.

Figura 15 – Distribuição de patentes por cluster com base em termos técnicos extraídos.



A Figura 15 mostra que alguns clusters concentraram maior número de patentes, como os clusters 2, 4 e 5, os quais merecerão atenção especial na análise temática. Essa variação pode refletir tanto a recorrência de certos vocabulários técnicos quanto a maior atividade de inovação em determinadas áreas.

Quadro-Resumo: Temas Centrais por Cluster

A Tabela 8 sintetiza os principais eixos temáticos identificados nos clusters formados a partir dos termos técnicos extraídos automaticamente. Essa visão consolidada permite reconhecer, de forma rápida, os focos tecnológicos predominantes em cada grupo. Observa-se a presença de áreas como agricultura, eletrônica, inteligência artificial, saúde e conforto humano, demonstrando a diversidade e a complexidade das inovações contidas no conjunto analisado. Além de facilitar a categorização, o quadro contribui para uma interpretação mais estratégica do espaço tecnológico representado pelas patentes agrupadas.

Tabela 8 – Resumo temático dos clusters gerados a partir dos termos técnicos

Cluster	Tema Central
0	Dispositivos eletrônicos, médicos e sistemas interativos.
1	Inovação agrícola e biodefensivos.
2	IA, automação e controle de equipamentos.
3	Vestíveis, conforto e cuidados pessoais.
4	Maquinário agrícola e tratamento do solo.
5	Controle de pragas e dispositivos diversos.
6	Manejo biológico e sistemas agrícolas/animais.
7	Dispositivos eletrônicos, IoT e processamento de dados.

Fonte: Elaborado pelo autor, 2025.

A análise dos temas centrais por cluster revela a diversidade tecnológica presente no conjunto de patentes processadas. Observa-se uma segmentação coerente, com agrupamentos que refletem diferentes domínios de especialização: enquanto alguns clusters evidenciam avanços na agricultura e biotecnologia (Clusters 1, 4 e 6), outros concentram-se em dispositivos médicos, conforto pessoal e vestíveis (Clusters 0 e 3). A presença de clusters voltados a sistemas inteligentes, como os Clusters 2 e 7, indica uma forte tendência de integração entre hardware, automação e inteligência artificial.

Esse agrupamento automatizado foi viabilizado a partir do modelo multidimensional previamente implementado, no qual os dados foram organizados segundo uma arquitetura em fatos e dimensões. Essa estrutura permitiu não apenas a associação eficiente entre os termos técnicos gerados por LLMs e suas respectivas patentes, como também viabilizou consultas complexas e contextualizadas, fundamentais para a identificação de padrões temáticos e a análise exploratória dos dados tecnológicos.

A visualização por clusters permite uma visão macro de como as inovações se distribuem tecnicamente, sendo uma ferramenta útil para:

- Identificar nichos tecnológicos com maior concentração de patentes;
- Detectar convergências entre áreas distintas (por exemplo, saúde e IoT);

- Direcionar análises prospectivas ou comparativas com base em vocabulário técnico comum.

A separação por cluster pode servir como base para análises mais aprofundadas, como geração de nuvens de palavras por agrupamento, mapeamento de termos emergentes e definição de estratégias para inovação tecnológica em setores específicos.

4.5 CONSUMO DE DADOS VIA API REST

A arquitetura desenvolvida contempla uma camada de exposição de dados por meio de uma API REST, projetada para facilitar o acesso estruturado às informações armazenadas no *Data Warehouse*. Essa interface permite que usuários e sistemas externos consultem estatísticas analíticas diretamente a partir do modelo dimensional, viabilizando o uso dos dados em aplicações diversas como dashboards, notebooks analíticos, sistemas de recomendação e plataformas de apoio à decisão.

A API foi implementada com rotas parametrizáveis, organizadas de forma intuitiva e documentadas automaticamente via Swagger UI, o que proporciona maior acessibilidade para desenvolvedores e analistas. Entre os principais recursos disponíveis, destacam-se:

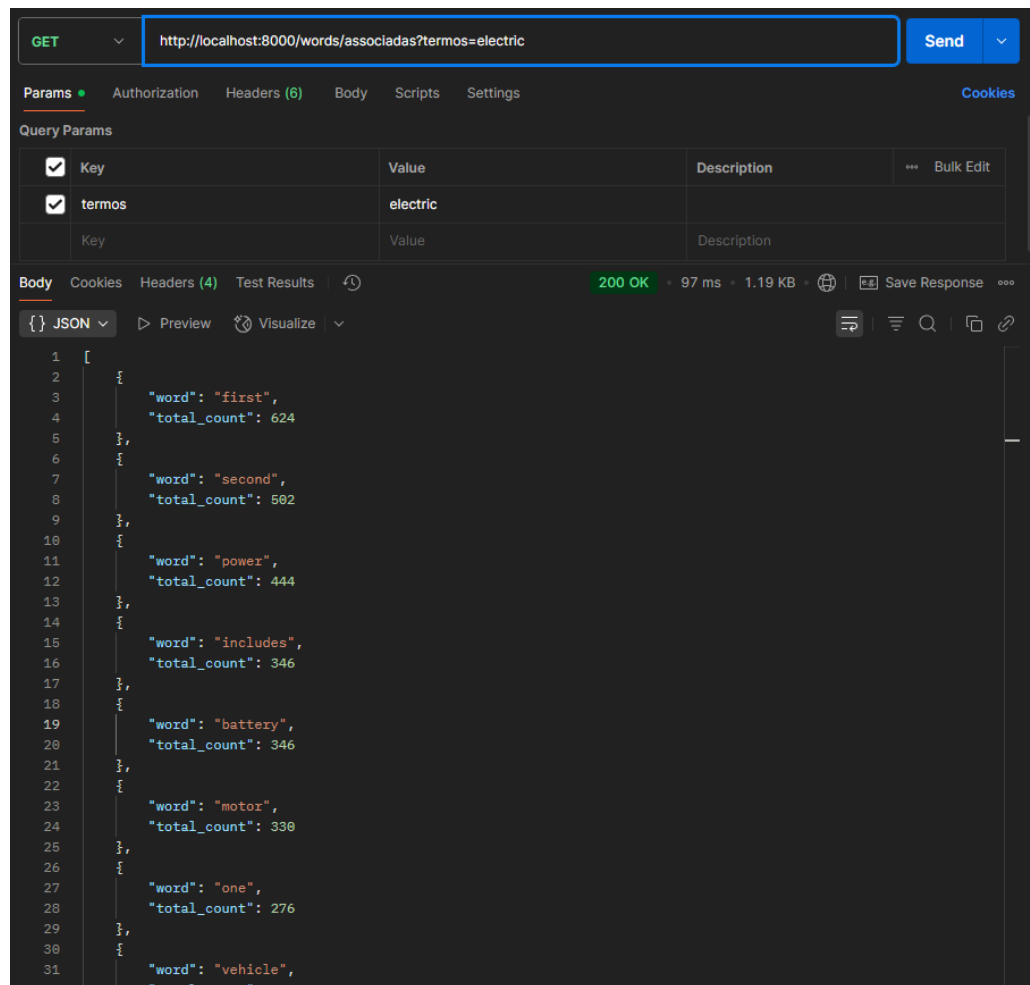
- Consulta de termos técnicos mais frequentes ao longo do tempo;
- Agrupamento de termos por país, autor, categoria ou período;
- Busca por entidades técnicas associadas com base em coocorrência semântica;
- Consulta temporal detalhada por termo específico;
- Filtragem de resultados por atributos multidimensionais do modelo, como year, subclass ou country.

Além de sua flexibilidade funcional, a API também se destacou em termos de desempenho. Testes realizados em ambiente local indicaram tempos de resposta médios inferiores a 80ms para a maioria dos endpoints, mesmo em consultas envolvendo agregações e junções entre múltiplas tabelas.

A Figura 16 ilustra um exemplo de uso da API via a ferramenta Postman. No exemplo apresentado, é feita uma requisição ao endpoint `/words/associadas` utilizando o termo `electric`, com o objetivo de recuperar palavras semanticamente relacionadas a esse conceito com base em sua coocorrência nos resumos das patentes. O retorno da requisição contém os termos associados e suas respectivas contagens de ocorrência conjunta, possibilitando a inferência de relações tecnológicas latentes na base de dados.

Essa camada de API amplia significativamente o potencial da arquitetura, permitindo que os dados analíticos sejam integrados de forma transparente a sistemas externos. Ela transforma o *Data Warehouse* em uma plataforma acessível e programável, pronta para uso em soluções baseadas em inteligência de dados, monitoramento tecnológico e visualizações dinâmicas personalizadas.

Figura 16 – Exemplo de consumo da API no endpoint /words/associadas utilizando o termo electric.



Fonte: Elaborado pelo autor, 2025.

4.6 ANÁLISE CRÍTICA DA ARQUITETURA

A arquitetura proposta demonstrou ser eficiente para a ingestão, organização e análise de dados textuais de patentes. Entre os principais pontos fortes destacam-se:

- Estrutura modular e reproduzível via Docker;
- Automação completa do processo de ETL;
- O processo de tokenização já inclui a remoção de *stopwords*, o que contribui para a redução de ruído textual. Futuramente, pode-se ainda incorporar técnicas de lematização ou stemming para melhorar a análise semântica;
- Modelo de dados dimensionais apropriado para consultas OLAP;
- Exposição de dados via API com endpoints dinâmicos e documentados;
- Flexibilidade para integração futura com modelos de IA;
- Implementação de extração de entidades técnicas por meio de LLMs, com armazenamento estruturado dos termos gerados e sua normalização no modelo dimensional, o que amplia

significativamente as possibilidades analíticas e semânticas da base.

A combinação entre técnicas tradicionais de pré-processamento e modelagem dimensional com abordagens modernas baseadas em LLMs proporciona uma camada semântica adicional ao sistema. A extração e normalização de entidades técnicas diretamente do texto das patentes permite realizar análises mais ricas e contextualizadas, como o rastreamento de tecnologias emergentes ao longo do tempo, a identificação de padrões de inovação por país ou autor, e a construção de painéis interativos mais expressivos. Essa sinergia amplia consideravelmente o valor do Data Warehouse como ferramenta de apoio à gestão do conhecimento técnico-científico.

4.7 DISCUSSÃO DOS RESULTADOS

Os resultados alcançados demonstram a viabilidade e os benefícios da integração de modelos de linguagem de última geração (LLMs) ao processo de ETL voltado à análise de patentes. A extração automatizada de entidades técnicas diretamente dos textos, como resumos e títulos, permitiu ultrapassar as limitações da contagem de palavras isoladas, aproximando a arquitetura de uma análise semântica mais sofisticada e contextualizada.

A incorporação desses termos técnicos ao modelo dimensional viabilizou consultas SQL simples, porém semanticamente ricas, sem depender de infraestrutura adicional ou motores de busca especializados. O processo de extração, normalização e persistência dos termos foi totalmente automatizado, garantindo reprodutibilidade e escalabilidade. Isso torna o sistema apto a lidar com bases em constante atualização, como é o caso dos repositórios semanais da USPTO.

A arquitetura de *Data Warehouse* proposta demonstrou-se eficaz, viável e tecnicamente replicável. A automação completa do *pipeline* ETL e a estruturação dos dados em um modelo dimensional facilitaram de forma significativa a análise de informações técnicas complexas presentes nas patentes. A exposição desses dados via API REST contribuiu para democratizar o acesso a conteúdos especializados, promovendo sua reutilização em sistemas analíticos, dashboards e plataformas de apoio à decisão.

Além disso, o uso de dados abertos, ferramentas livres (Python, PostgreSQL, Docker) e boas práticas de engenharia de dados comprova o potencial de soluções de baixo custo aplicadas à inovação tecnológica baseada em informação. A integração entre ciência de dados, gestão do conhecimento e fontes públicas de informação tecnológica representa um avanço concreto na construção de soluções inteligentes e orientadas por dados.

4.8 LIMITAÇÕES E PERSPECTIVAS FUTURAS

Embora os resultados obtidos com a arquitetura desenvolvida tenham demonstrado a viabilidade técnica e analítica da solução, algumas limitações ainda persistem e abrem caminho para futuras melhorias.

Em primeiro lugar, os dados utilizados neste trabalho foram extraídos de apenas dois arquivos XML da base USPTO, correspondentes a semanas específicas de publicação. Apesar de

essa amostra ter permitido a validação da arquitetura, a adoção de um volume maior de arquivos, com ingestão contínua dos arquivos semanais, possibilitaria análises mais robustas, além de testar a escalabilidade do pipeline em cenários com maior carga de dados e complexidade.

Em relação à extração de entidades técnicas com apoio de LLMs, observou-se grande valor na aplicação prática da abordagem, incluindo a normalização e integração dos termos ao modelo dimensional. No entanto, ainda há espaço para aprimoramentos, como a adoção de técnicas complementares de pré-processamento, como lematização, stemming ou agrupamento de termos semanticamente equivalentes, de forma a reduzir redundâncias e melhorar a coesão dos dados extraídos.

Outro aspecto a ser considerado refere-se à validação qualitativa dos termos gerados. Embora as entidades extraídas tenham se mostrado relevantes em análises exploratórias, não foi realizada uma validação manual sistemática por especialistas ou a aplicação de métricas específicas de qualidade semântica, o que seria desejável em uma futura etapa.

No plano da arquitetura, também não foram implementados mecanismos automatizados de versionamento de dados, controle de qualidade ou auditoria das transformações aplicadas no pipeline, funcionalidades importantes para ambientes produtivos com requisitos de rastreabilidade e governança de dados.

Por fim, ainda que alguns experimentos de aplicação de IA tenham sido conduzidos, como a regressão para previsão de ocorrências futuras e o agrupamento (clustering) de patentes por perfil técnico, existe espaço para aprofundamento dessas aplicações. Modelos de classificação de patentes por categoria tecnológica, sistemas de recomendação de tecnologias semelhantes e interfaces interativas com dashboards enriquecidos são evoluções viáveis e coerentes com a base construída.

Como perspectivas futuras, destaca-se a expansão do modelo dimensional para incluir novos atributos relevantes (como coautores, empresas, classificação CPC e citações), a ingestão contínua e escalável de dados da USPTO, o refinamento das técnicas de extração de entidades e a integração com ferramentas de visualização interativa. Essas melhorias poderão transformar a solução desenvolvida em uma plataforma poderosa de análise tecnológica, prospecção de inovação e apoio à gestão do conhecimento em ambientes reais de pesquisa, desenvolvimento e inteligência competitiva.

Considerações Finais

De forma geral, os resultados apresentados demonstram que a arquitetura proposta atendeu aos objetivos estabelecidos no início deste trabalho. Foi possível construir uma solução automatizada e modular para ingestão, transformação, armazenamento e análise de dados textuais de patentes, integrando técnicas de engenharia de dados com recursos avançados de PLN. A adoção de LLMs para a extração de entidades técnicas ampliou significativamente a profundidade das análises possíveis, agregando uma camada semântica valiosa ao modelo dimensional. A arquitetura mostrou-se eficaz tanto do ponto de vista técnico quanto conceitual, reforçando o

potencial dos *Data Warehouses* como plataformas estratégicas para a gestão do conhecimento em contextos tecnológicos.

5 CONCLUSÃO

Este trabalho apresentou uma arquitetura de *Data Warehouse* automatizado voltada ao processamento, organização e análise de dados textuais extraídos de documentos de patentes. A proposta integrou fundamentos clássicos da engenharia de dados com técnicas modernas de PLN, com o objetivo de apoiar a gestão do conhecimento tecnológico de forma mais eficiente, estruturada e escalável.

Os objetivos inicialmente definidos foram plenamente alcançados. Foi implementado um *pipeline* ETL automatizado capaz de extrair dados relevantes de arquivos XML da base USPTO, realizar transformações com limpeza e tokenização textual, e carregar essas informações em um modelo dimensional otimizado para consultas analíticas. A disponibilização dos dados via API REST completou a arquitetura, permitindo sua integração com sistemas externos e democratizando o acesso à informação estratégica.

Um dos principais diferenciais do sistema proposto foi a incorporação de modelos de linguagem de última geração (LLMs), como o GPT-4o, para a extração automática de termos técnicos relevantes dos resumos das patentes. Esse enriquecimento semântico elevou a profundidade das análises possíveis, viabilizando o rastreamento de tecnologias emergentes, a identificação de padrões temáticos e a geração de indicadores úteis à prospecção e inteligência tecnológica. A integração desses termos ao modelo dimensional, com normalização e inserção em tabelas fato, demonstrou-se tecnicamente robusta e conceitualmente valiosa.

Apesar dos resultados promissores, algumas limitações devem ser consideradas. Os dados analisados neste estudo derivam de apenas dois arquivos XML da USPTO, representando uma fração reduzida do volume total publicado semanalmente. Além disso, aspectos como a extração de entidades com classes semânticas específicas, a desambiguação de termos compostos e a exploração semântica da descrição completa das patentes ainda oferecem espaço para aprimoramentos.

Como perspectivas futuras, destaca-se a expansão da base de patentes processadas, o aprimoramento das técnicas de normalização e agregação semântica, o uso de modelos de IA locais que garantam maior autonomia operacional e a criação de painéis interativos para visualização exploratória dos resultados. Vislumbra-se também a integração da solução com sistemas de recomendação, módulos de monitoramento tecnológico e ferramentas de apoio à inovação.

Conclui-se, portanto, que a arquitetura desenvolvida representa uma contribuição prática, eficiente e inovadora para o uso de *Data Warehouses* no contexto da análise tecnológica. Ao integrar engenharia de dados, PLN e modelos generativos, a solução proposta abre caminhos concretos para aplicações em inteligência competitiva, gestão do conhecimento e inovação orientada por dados.

REFERÊNCIAS

ALMEIDA, Eduardo Cunha de. Estudo de viabilidade de uma plataforma de baixo custo para data warehouse. **arXiv preprint arXiv:1108.0729**, 2011. Acesso em 14 jun. 2025. Disponível em: <https://arxiv.org/abs/1108.0729>. Citado na página 23.

DEVLIN, Jacob et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In: **Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)**. [s.n.], 2019. p. 4171–4186. Acesso em: 10 jun. 2025. Disponível em: <https://aclanthology.org/N19-1423.pdf>. Citado na página 20.

FALL, Charles J et al. Automated categorization of patent documents. In: ACM. **Proceedings of the ACM Symposium on Document Engineering**. [S.l.], 2003. p. 93–100. Citado na página 24.

GOLFARELLI, Matteo; RIZZI, Stefano. **Data warehouse design: Modern principles and methodologies**. McGraw-Hill, Inc., 2009. Acesso em: 14 jun. 2025. Disponível em: <https://dl.acm.org/doi/abs/10.5555/1594749>. Citado na página 17.

GONZÁLEZ, S. M.; SAKATA, T. C.; NOGUEIRA, R. R. Newsminer: Enriched multidimensional corpus for text-based applications. In: **Artificial Intelligence and Soft Computing: ICAISC 2020**. Springer, 2020. p. 231–242. Acesso em: 20 mai. 2025. Disponível em: https://link.springer.com/chapter/10.1007/978-3-030-61534-5_21. Citado 2 vezes nas páginas 10 e 13.

INMON, William H. **Building the data warehouse**. [S.l.]: John Wiley & sons, 2005. Citado 3 vezes nas páginas 9, 15 e 20.

Instituto Nacional da Propriedade Industrial (INPI). **Guia Básico de Patentes**. 2025. Acesso em: 15 jun. 2025. Disponível em: <https://www.gov.br/inpi/pt-br/servicos/patentes/guia-basico>. Citado na página 14.

JOULIN, Armand et al. Bag of tricks for efficient text classification. **arXiv preprint arXiv:1607.01759**, 2016. Acesso em: 15 mai. 2025. Disponível em: <https://arxiv.org/pdf/1607.01759>. Citado na página 20.

JUNIOR, Vanderlei FREITAS et al. A pesquisa científica e tecnológica. **Revista ESPACIOS| Vol. 35 (Nº 09) Año 2014**, 2014. Acesso em: 14 jun. 2025. Disponível em: <http://sistemasblandosxd.revistaespacios.com/a14v35n09/14350913.html>. Citado na página 12.

KIMBALL, Ralph; ROSS, Margy. **The data warehouse toolkit: The definitive guide to dimensional modeling**. [S.l.]: John Wiley & Sons, 2013. Citado 4 vezes nas páginas 9, 12, 16 e 18.

KRESTEL, Ralf et al. A survey on deep learning for patent analysis. **World Patent Information**, Elsevier, v. 65, p. 102035, 2021. Acesso em 16 jun. 2025. Disponível em: https://www.researchgate.net/profile/Ralf-Krestel/publication/350523275_A_survey_on_deep_learning_for_patent_analysis/links/6093f15e92851c490fbc9660/A-survey-on-deep-learning-for-patent-analysis.pdf. Citado na página 9.

LEE, Jinhyuk; KANG, Jisu; KANG, Jaewoo. A survey on deep learning for patent analysis. **Information Processing & Management**, v. 59, n. 6, p. 103064, 2022. Citado na página 23.

LIU, Qi; KUSNER, Matt J; BLUNSOM, Phil. A survey on contextual embeddings. **arXiv preprint arXiv:2003.07278**, 2020. Citado na página 22.

MACHADO, Gustavo V et al. Dod-etl: distributed on-demand etl for near real-time business intelligence. **Journal of Internet Services and Applications**, Springer, v. 10, n. 1, p. 21, 2019. Acesso em 16 jun. 2025. Disponível em: <https://link.springer.com/content/pdf/10.1186/s13174-019-0121-z.pdf>. Citado na página 23.

MIHALCEA, Rada; TARAU, Paul. Textrank: Bringing order into texts. **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, p. 404–411, 2004. Acesso em 15 jun. 2025. Disponível em: <https://aclanthology.org/W04-3252.pdf>. Citado 2 vezes nas páginas 19 e 24.

MIKOLOV, Tomas et al. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013. Acesso em: 15 mai. 2025. Disponível em: <https://arxiv.org/pdf/1301.3781>. Citado 2 vezes nas páginas 20 e 21.

NADEAU, David; SEKINE, Satoshi. A survey of named entity recognition and classification. **Linguisticae Investigationes**, John Benjamins, v. 30, n. 1, p. 3–26, 2007. Acesso em 18 mai. 2025. Disponível em: <https://www.time.mk/trajkovski/thesis/li07.pdf>. Citado na página 19.

NOGUEIRA, Rodrigo Ramos. **Business Intelligence na prática: Modelagem Multidimensional e Data Warehouse**. 1. ed. Indaial: UNIASSELVI, 2020. ISBN 978-85-515-0454-3. Citado na página 16.

PENG, Jianjun et al. A knowledge graph-based method for patent analysis: A case of artificial intelligence technology. **World Patent Information**, v. 61, p. 101964, 2020. Citado 2 vezes nas páginas 8 e 11.

PENNINGTON, Jeffrey; SOCHER, Richard; MANNING, Christopher D. Glove: Global vectors for word representation. In: **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1532–1543. Citado na página 20.

RADFORD, Alec et al. Improving language understanding by generative pre-training. San Francisco, CA, USA, 2018. Acesso em 10 jun. 2025. Disponível em: <https://www.mikecaptain.com/resources/pdf/GPT-1.pdf>. Citado na página 20.

SILVA, João Victor Fernandes de Souza. Comparação de desempenho entre os bancos de dados postgresql e neo4j para acesso a dados complexos. Universidade Federal de Uberlândia, 2023. Acesso em 15 jun. 2025. Disponível em: <https://repositorio.ufu.br/bitstream/123456789/38755/1/ComparacaoDesempenhoEntre.pdf>. Citado na página 23.

TIKK, Domonkos et al. A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. **PLoS computational biology**, Public Library of Science San Francisco, USA, v. 6, n. 7, p. e1000837, 2010. Acesso em 11 jun. 2025. Disponível em: <https://journals.plos.org/ploscompbiol/article/file?id=10.1371/journal.pcbi.1000837&type=printable>. Citado na página 19.

TURNEY, Peter D. Learning algorithms for keyphrase extraction. In: **Information Retrieval**. [s.n.], 2000. v. 2, n. 4, p. 303–336. Acesso em 25 mai. 2025. Disponível em: <https://link.springer.com/content/pdf/10.1023/A:1009976227802.pdf>. Citado 2 vezes nas páginas 19 e 24.

WANG, Wei et al. **EvoPat: A Multi-Agent System for Interpretable Patent Analysis Using LLMs**. 2024. ArXiv preprint arXiv:2412.18100. Acesso em: 15 jun. 2025. Disponível em: <https://arxiv.org/abs/2412.18100>. Citado na página 23.

World Intellectual Property Organization (WIPO). **The International Patent System (PCT)**. 2025. Acesso em: 14 jun. 2025. Disponível em: <https://www.wipo.int/pct/en/>. Citado na página 14.

XU, Hongzhi et al. Big data cleaning: Problems and current approaches. **IEEE Access**, IEEE, v. 4, p. 1033–1043, 2016. Citado na página 9.

YOON, Byungun; PARK, Youngtae; KIM, Kunwoo. Detecting emerging keywords for technology trend analysis. **Technological Forecasting and Social Change**, Elsevier, v. 70, n. 2, p. 171–187, 2004. Citado na página 24.

ZHANG, Simona. Data modeling and etl with postgresql. 2020. Citado na página 23.