UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

CENTRO DE CIÊNCIAS EXATAS, NATURAIS E DA SAÚDE

DEPARTAMENTO DE COMPUTAÇÃO

MARIANA CUNHA DE OLIVEIRA

UMA ARQUITETURA DE *BUSINESS INTELLIGENCE* PARA INTEGRAÇÃO DE DADOS DE PESQUISAS NO BRASIL

ALEGRE - ES

2026

UMA ARQUITETURA DE *BUSINESS INTELLIGENCE* PARA INTEGRAÇÃO DE DADOS DE PESQUISAS NO BRASIL

Trabalho de conclusão de curso apresentado ao Departamento de Computação do Centro de Ciências Exatas, Naturais e da Saúde da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

por

MARIANA CUNHA DE OLIVEIRA

Orientador

ANTONIO ALMEIDA DE BARROS JUNIOR

Universidade Federal do Espírito Santo

ALEGRE - ES

2026

Mariana Cunha de Oliveira

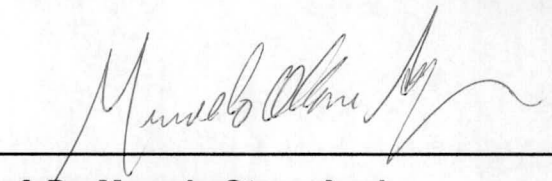
Uma arquitetura de *Business Intelligence* para integração de dados de pesquisas no Brasil

Trabalho de Conclusão de Curso apresentado ao Departamento de Computação do Centro de Ciências Exatas, Naturais e da Saúde da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

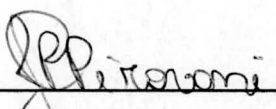
Aprovado em 24 de fevereiro de 2026.

COMISSÃO EXAMINADORA

Prof. Dr. Antonio Almeida de Barros Junior
Universidade Federal do Espírito Santo
Orientador



Prof. Dr. Marcelo Otone Aguiar
Universidade Federal do Espírito Santo



Prof.ª Drª. Juliana Pinheiro Campos Pirovani
Universidade Federal do Espírito Santo

AGRADECIMENTOS

Agradeço primeiramente a Deus, por ter me concedido forças ao longo de toda essa trajetória e por ter me sustentado nos momentos mais desafiadores, permitindo que eu chegasse até aqui.

À minha família, meu mais sincero agradecimento pelo apoio incondicional. Em especial, aos meus pais, Edna e Evânio, que trabalharam incansavelmente para tornar possível a realização deste sonho, sempre me incentivando e acreditando no meu potencial. À minha irmã, Larissa, agradeço por ser um exemplo constante de força, resiliência e inspiração.

Ao meu companheiro de vida, Rafael, deixo meu profundo agradecimento por nunca soltar a minha mão, pelo apoio, incentivo e compreensão em todos os momentos. Sua presença, confiança e incentivo tornaram essa caminhada mais leve e possível.

À minha prima Jéssica, que já estudava no *campus* de Alegre e me incentivou a dar esse primeiro passo, meu muito obrigada. Seu incentivo foi essencial e marcou o início dessa jornada.

Aos demais familiares, tios, avô, primos e amigos da família, agradeço pelo carinho, acolhimento e preocupação ao longo desse percurso. Mesmo à distância, cada gesto de apoio, especialmente nos momentos em que fui recebida com tanto carinho na Bahia, durante os períodos de férias, contribuiu para que eu seguisse em frente.

Ao meu orientador, Antonio, agradeço pelos ensinamentos, pela orientação e pelo suporte oferecido ao longo da graduação e na elaboração deste trabalho.

A todos os meus professores, minha gratidão pelos conhecimentos transmitidos. Iniciei essa trajetória com muitos desafios e hoje encerro este ciclo com um enorme aprendizado acadêmico, profissional e pessoal.

Aos amigos e colegas que fiz na Universidade Federal do Espírito Santo, companheiros de curso e de trabalhos acadêmicos, agradeço pela convivência, pelas trocas de conhecimento e pelos momentos compartilhados ao longo dessa caminhada.

Por fim, agradeço à Universidade Federal do Espírito Santo (UFES) pela oportunidade de formação e por proporcionar um ambiente de aprendizado que contribuiu significativamente para meu crescimento.

RESUMO

A produção científica de professores universitários é um indicador fundamental para avaliar o avanço da ciência no Brasil. Monitorar essa produção permite compreender a distribuição do conhecimento acadêmico no país, identificar áreas estratégicas e direcionar políticas públicas que incentivem a pesquisa e o desenvolvimento. Desta forma, este trabalho propôs a construção de uma solução de *Business Intelligence* [1](BI), baseada em dados abertos para mapear e analisar a produção científica dos docentes universitários no Brasil. A metodologia adotada consiste na utilização de arquivos XML dos currículos da Plataforma Lattes, os quais foram transformados para o formato JSON, mantendo apenas os campos de interesse para a análise. Os dados foram armazenados em um *Data Warehouse*, o que permite a construção de consulta por gestores institucionais. Para visualização dos resultados foram elaborados *dashboards,* construídos no *Streamlit*, uma ferramenta que possibilita a construção de interfaces interativas para análise de dados, facilitando a interpretação das informações e possibilitando a formulação de estratégias para fortalecer o desenvolvimento científico e tecnológico no Brasil. Como resultados, a solução desenvolvida possibilitou a consolidação e organização estruturada dos dados de produção científica, permitindo análises dinâmicas por instituição, área de conhecimento, tipo de produção e período temporal. Os *dashboards* gerados evidenciaram padrões de concentração da produção científica, tendências de crescimento ao longo dos anos e diferenças entre áreas estratégicas, demonstrando o potencial da ferramenta como apoio à tomada de decisão por gestores acadêmicos e formuladores de políticas públicas.

Palavras-chave: produção científica; pesquisa; *Business Intelligence; Data Warehouse.*

[1] https://www.oracle.com/br/what-is-business-intelligence/

SUMÁRIO

1. INTRODUÇÃO 10
1.1 PROBLEMA E SUA IMPORTÂNCIA 11
1.2 OBJETIVOS 12
1.2.1 Objetivo geral 12
1.2.2 Objetivos específicos 13
2. REVISÃO DE LITERATURA 14
2.1 TECNOLOGIA APLICADA 14
2.1.1 TRABALHO DE ANDRADE (2017) 14
2.2 TRABALHOS RELACIONADOS 15
2.2.1 TRABALHO DE MACHADO E MEIRELLES (2005) 17
2.2.2 TRABALHO DE ANDRADE ET AL. (2010) 20
2.2.3 TRABALHO DE FERREIRA E SILVA JÚNIOR (2016) 21
2.2.4 TRABALHO DE ESTÁCIO (2017) 24
3. METODOLOGIA 26
3.1 FERRAMENTAS UTILIZADAS 28
3.1.1 Python 28
3.1.2 PostgreSql 28
3.1.4 Streamlit 29
3.1.5 Base de dados de origem 29
3.1.6 Dbeaver 30
4. RESULTADOS OBTIDOS 32
4.1 LEVANTAMENTO DOS INDICADORES 32

4.2 ESTRUTURA DO PROJETO 33
4.2.1 Pasta db 33
4.2.2 Pasta lattes_tcc 34
4.2.3 Pasta stage 34
4.2.4 Pasta sql_criar_tabelas 34
4.2.5 Pasta populando_tabelas 34
4.2.6 Pasta streamlit 35
4.3 MODELO DIMENSIONAL 35
4.3.1 dim_pesquisador 36
4.3.2 dim_area 36
4.3.3 dim_linha_pesquisa 36
4.3.4 dim_tempo 36
4.3.5 dim_tipo_producao 36
4.3.6 dim_localizacao_trabalhos 36
4.3.7 fato_pesquisador_producoes 36
4.3.8 fato_pesquisador_area_atuacao 37
4.3.9 fato_pesquisador_linha_pesquisa 37
4.3.10 fato_pesquisador_producao_localizacao 37
4.4 GRÁFICOS 38
5. CONCLUSÕES 45
6. REFERÊNCIAS 47
APÊNDICE A - DICIONÁRIO DE DADOS 50

LISTA DE FIGURAS

Figura 1 - Evolução do número de currículos registrados na base do Currículo Lattes. 25
Figura 2 - Arquitetura de BI proposta. 26
Figura 3 - Estrutura do projeto no Github. 33
Figura 4 - Modelo Dimensional. 35
Figura 5 - Estrutura do banco de dados visualizada no DBeaver, destacando os esquemas DW e STG. 38
Figura 6 - Análise da distribuição de Pesquisadores por Grande Área. 40
Figura 7 - Ranking das 20 linhas de pesquisa com maior número de pesquisadores. 41
Figura 8 - Distribuição por tipo de produção. 42
Figura 9 - Evolução anual do número de pesquisadores ativos. 43
Figura 10 - Quantidade de apresentações de trabalho que ocorreram na UFES. 44

LISTA DE TABELAS

Tabela 1 - Resultados da busca na plataforma Google Scholar para palavras-chaves em português. 16
Tabela 2 - Resultados da busca na plataforma Google Scholar para palavras-chave em inglês. 16
Tabela 3 - Produção por Titulação dos Docentes da Área de Filosofia e Ciências Humanas/UFBA, Período de 1995 a 1999. 18
Tabela 4 -Identificação dos Jovens-Doutores com bolsa Produtividade em Pesquisa, por códigos de identificação, ano de doutorado, quantidade de pós-doutoramento e ingresso na UFSCar. 22

1. INTRODUÇÃO

A produção científica no Brasil apresenta variações significativas entre diferentes áreas do conhecimento. Algumas áreas possuem maior volume de publicações devido a fatores como disponibilidade de financiamento, tradição acadêmica e estrutura de pesquisa consolidada.

Segundo dados da *Clarivate Analytics*, o Brasil publicou mais de 250 mil artigos na base *Web of Science* entre 2011 e 2016 (NETTO, 2019), abrangendo todas as áreas do conhecimento e posicionando-se na décima terceira colocação entre mais de 190 países em termos de produção científica (UFPA, 2019). Estudos mais recentes confirmam a continuidade e expansão dessa produção. Para se ter uma ideia, entre os anos de 2019 e 2023, o país registrou 6.304 publicações científicas sobre inteligência artificial, estando entre os 20 países com maior volume de pesquisa nessa área, conforme relatório divulgado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) em parceria com a *Clarivate Analytics* (CAPES, 2024). Esse avanço reflete o fortalecimento das universidades e centros de pesquisa brasileiros, bem como o aumento do investimento e da diversificação das áreas de estudo, especialmente em campos emergentes como a inteligência artificial.

Apesar do crescimento da produção científica no Brasil, há desafios estruturais que impactam diretamente o seu desenvolvimento e disseminação. A escassez de financiamento contínuo, a concentração de investimentos em determinadas áreas e regiões, a dependência de periódicos internacionais para publicação e a dificuldade de acesso aberto a muitos trabalhos. Além disso, questões como burocracia para captação de recursos, fuga de talentos para o exterior e desigualdade na distribuição da infraestrutura de pesquisa, dificultam ainda mais o desenvolvimento de um ambiente científico integrado no país. Todos esses são obstáculos que limitam o alcance e a influência da ciência nacional.

Além dos desafios já citados, os órgãos de fomento, as instituições de ensino e os agentes governamentais, também encontram dificuldades para monitorar a produtividade das pesquisas desenvolvidas. No processo de monitoramento, a dispersão dos dados e a falta de informações estruturadas e integradas são certamente os principais desafios. Sabe-se que, a análise de dados pode ser

dificultada quando há uma fragmentação das informações, o que impossibilita sua consolidação e interpretação. Em instituições de ensino superior, especificamente nos casos das universidades federais, onde concentram-se a maior parte das pesquisas no país, uma ferramenta que integre os dados de pesquisa pode auxiliar gestores no processo de tomada de decisão.

Desta forma, estudos sobre a produção científica em instituições de ensino superior são importantes, visto que envolve recursos e bolsas. Tais recursos muitas vezes são distribuídos sem uma análise detalhada e de forma justa. Para se ter uma ideia, pesquisas como de Reis, et al. (2008), apontam que as condições de trabalho dos docentes, a infraestrutura das universidades, a intensificação do trabalho e a cobrança por produtividade acadêmica, são fatores que contribuem para o adoecimento dos professores nas universidades públicas. Para Gemelli e Gloss (2022), a precarização e a intensificação do trabalho docente são temas amplamente discutidos na literatura acadêmica, e destacam o impacto direto que essas questões exercem sobre a produção científica.

Vale ressaltar também que, a ampliação das jornadas de trabalho, a desvalorização profissional e a cultura de avaliação baseada na produtividade quantitativa, criam um ambiente desfavorável ao desenvolvimento de pesquisas de longo prazo e de alta qualidade. Portanto, o desenvolvimento de ferramentas tecnológicas, que facilitem a organização, consolidação e análise dos dados da produção científica, pode ser um aliado à gestão neste processo, especialmente quanto à forma de distribuição dos recursos de pesquisa. A partir da consolidação desses tipos de informação, é possível que a instituição avalie e incentive áreas que ainda não desenvolvem projetos de pesquisas.

Neste contexto, o presente estudo visa a construção de uma arquitetura de *Business Intelligence* (BI), baseada na modelagem dimensional de dados provenientes dos currículos da Plataforma *Lattes*.

1.1 PROBLEMA E SUA IMPORTÂNCIA

A plataforma *Lattes*[2] constitui a principal base nacional de informações sobre

[2] https://lattes.cnpq.br/

pesquisadores brasileiros, reunindo dados detalhados sobre formação acadêmica, trajetória profissional e produção científica. Sua ampla utilização no meio acadêmico reforça sua importância como fonte de dados, para análises relacionadas à pesquisa científica e à atuação docente nas instituições de ensino superior.

Entretanto, apesar de sua abrangência e relevância, a utilização dos dados do Currículo *Lattes* para fins analíticos, apresenta limitações significativas. As informações são disponibilizadas de forma altamente dispersa e em formatos que não foram concebidos para análises gerenciais, como o formato *XML*, o que dificulta a integração, a consolidação e a interpretação dos dados. A fragmentação das informações pode comprometer a análise da produção científica, resultando em interpretações parciais ou inconsistentes.

Além das dificuldades relacionadas à estrutura dos dados, o acesso a grandes volumes de informações do Currículo *Lattes* envolve restrições técnicas e institucionais. A obtenção de dados em larga escala requer autorização formal, vínculo com instituição de ensino ou pesquisa e o atendimento a requisitos específicos de infraestrutura. Essas condições tornam o processo de coleta complexo e limitam iniciativas voltadas à análise abrangente da produção científica.

Dessa forma, o problema central deste trabalho consiste na dificuldade de acesso, organização e análise estruturada dos dados de produção científica, o que impacta diretamente a capacidade de instituições e gestores acadêmicos acompanharem a evolução das pesquisas, identificarem padrões e utilizarem essas informações de forma estratégica.

Com o objetivo de mitigar esse problema, este trabalho propõe o desenvolvimento de uma arquitetura de *Business Intelligence* baseada nos dados da plataforma *Lattes*, obtidos por meio de acesso institucional autorizado.

A partir da abordagem proposta, espera-se oferecer uma solução inovadora e eficiente como facilitadora no processo de consulta e análise de dados relacionados às pesquisas desenvolvidas nas universidades.

1.2 OBJETIVOS

1.2.1 Objetivo geral

Este trabalho tem como objetivo geral desenvolver um ambiente de *Business Intelligence* que contenha um *Data Warehouse* (DW), com dados de pesquisas integrados, e que possibilite a exploração estruturada de informações sobre programas de pós-graduação, docentes e produção científica. O projeto contempla a construção de um modelo multidimensional baseado na abordagem de Kimball e Ross (2013), bem como a implementação de um processo de ETL (Extração, Transformação e Carga) e *dashboards* interativos que permitem a visualização de indicadores estratégicos e a extração de *insights*.

1.2.2 Objetivos específicos

OE1. Realizar o levantamento e a compreensão dos dados dos currículos da Plataforma *Lattes*, identificando informações relevantes sobre docentes e produção científica;

OE2. Identificar os indicadores em consonância com os dados obtidos;

OE3. Implementar um ambiente de *Data Warehouse*;

OE4. Implementar o processo de ETL (Extração, Transformação e Carga) que integre e prepare os dados para o armazenamento no *Data Warehouse*;

OE5. Construir *dashboards* que reflitam os indicadores definidos e que possibilitem a análise dos dados integrados sobre produção científica.

2. REVISÃO DE LITERATURA

A revisão bibliográfica deste trabalho encontra-se organizada em duas seções. Na primeira, são abordados os principais conceitos e fundamentos da tecnologia de *Data Warehouse* (DW), tomando como referência o estudo de Andrade (2017), destacando sua importância para a integração, organização e análise de dados no contexto da gestão da informação. Na segunda, são discutidos estudos relacionados à produção científica de pesquisadores, bem como análises fundamentadas em dados da Plataforma *Lattes*, evidenciando sua relevância para a construção e interpretação de indicadores acadêmicos.

2.1 TECNOLOGIA APLICADA

2.1.1 TRABALHO DE ANDRADE (2017)

O *Business Intelligence* (BI) pode ser compreendido como um conjunto de metodologias, processos e tecnologias voltadas à coleta, organização, análise e disponibilização de dados com o objetivo de apoiar a tomada de decisão nas organizações. Conforme discutido no e-book BI e *Data Warehousing* na Prática, o BI não se limita à construção de relatórios ou *dashboards*, mas envolve uma arquitetura estruturada que transforma dados operacionais em informações estratégicas. Nesse contexto, o *Data Warehouse* (DW) assume papel central, funcionando como um repositório integrado e orientado à análise.

O *Data Warehouse* é definido como um banco de dados projetado especificamente para consultas analíticas e suporte à decisão. Diferentemente dos sistemas transacionais, que registram operações do dia a dia e priorizam desempenho em inserções e atualizações, o DW é estruturado para armazenar dados históricos, consolidados e organizados por assunto. Essa característica permite que gestores e analistas realizem consultas complexas, comparações temporais e análises estratégicas com maior eficiência.

Para que os dados sejam corretamente estruturados no *Data Warehouse*, é necessário que passem por um processo conhecido como ETL (Extração, Transformação e Carga). Esse processo é responsável por extrair dados de diferentes fontes, transformá-los conforme regras de negócio e padrões definidos, e carregá-los no ambiente analítico. A etapa de transformação é especialmente relevante, pois

envolve atividades como limpeza de dados, padronização de formatos, tratamento de inconsistências e adequação às estruturas do DW. Andrade(2017) destaca que a qualidade do processo de ETL impacta diretamente na qualidade das análises geradas, tornando essa etapa fundamental em projetos de BI.

No contexto da modelagem de dados, a modelagem dimensional é apresentada como a abordagem mais indicada para projetos *de Data Warehousing*. Ela organiza os dados em tabelas fato e tabelas dimensão. As tabelas fato concentram as métricas quantitativas do negócio, enquanto as tabelas dimensão armazenam atributos descritivos que contextualizam essas métricas, como tempo, categorias ou entidades relacionadas ao processo analisado. Essa estrutura favorece desempenho em consultas e facilita a interpretação dos dados pelos usuários.

A modelagem dimensional é amplamente adotada como base conceitual para projetos de *Data Warehouse*, por oferecer uma estrutura orientada à análise e ao desempenho em consultas. Essa metodologia propõe a organização incremental do *Data Warehouse* a partir de *Data Marts* estruturados por área de negócio, permitindo que a solução seja construída de forma evolutiva e alinhada às demandas organizacionais. Tal abordagem favorece a flexibilidade e a escalabilidade da arquitetura, além de reforçar a importância de um planejamento estruturado, no qual o modelo de dados seja concebido para facilitar tanto a manutenção quanto a expansão futura da solução.

Por fim, o autor enfatiza que a arquitetura de *Business Intelligence* é composta por diferentes camadas interdependentes, incluindo as fontes de dados, a área de *staging*, o processo de ETL, o *Data Warehouse* e a camada de visualização. Essa organização em camadas possibilita maior controle sobre o fluxo de dados, promove padronização e contribui para a governança da informação.

2.2 TRABALHOS RELACIONADOS

A revisão de literatura foi realizada inicialmente, por meio de uma busca na plataforma *Google Scholar*[3], utilizando como parâmetros de busca as palavras-chave: produção científica, docentes, pós-graduação, educação, currículo *lattes*. E em inglês: *“scientific production”*, *“business intelligence”*, *“teachers”* e *“data warehouse”*, bem

[3] https://scholar.google.com/

como a combinação desses termos. Vale ressaltar que a busca das palavras-chave foi conduzida de forma independente.

Para a seleção dos trabalhos encontrados, foi realizada uma análise preliminar baseada na leitura dos títulos, a fim de identificar aqueles mais alinhados ao tema proposto. Em seguida, os estudos considerados relevantes foram submetidos a uma leitura mais aprofundada, para uma análise detalhada de correlação com o objeto de estudo. Na Tabela 1, são apresentados seis trabalhos selecionados a partir da busca, sendo dois resultantes da combinação das palavras-chave “produção científica” e “docentes”, outros dois da combinação “produção científica” e “professores” e o restante da combinação “currículo lattes”.

Tabela 1 – Resultados da busca na plataforma *Google Scholar* para palavras-chaves em português.

Palavra-chave	**Resultados**	**Trabalhos selecionados**
Produção científica	2.490.000	0
Produção científica e docentes	840.000	2
Produção científica de professores	861.000	2
Docentes	5.630.000	0
Pós-graduação	1.470.000	0
Educação	4.510.000	0
Curriculo Lattes	19.600	2
Total	**15.801.000**	**6**

Fonte: O autor.

Na Tabela 2, são apresentados dois trabalhos pesquisados por termos em inglês, sendo que, um deles foi encontrado a partir da combinação das palavras-chave *“business” e “intelligence”* e o outro resultou na combinação das palavras-chave *“data” e “warehouse”*. Para essas combinações, foram encontrados dois trabalhos, porém os mesmos não foram selecionados por não se tratar do tema introdutório em si desta pesquisa.

Tabela 2 – Resultados da busca na plataforma *Google Scholar* para palavras-chave em inglês.

Palavra-chave	**Resultados**	**Trabalhos selecionados**
Scientific Production	6.470.000	0
Business Intelligence	4.500.000	1

Scientific Production and Business Intelligence	2.970.000	0
Teachers	5.400.000	0
Data Warehouse	1.270.000	1
Total	**20.610.000**	**2**

Fonte: O autor.

Dos dois documentos selecionados na Tabela 02, estes já haviam sido identificados em pesquisas anteriores realizadas por meio de buscas no navegador *Google* sobre o tema, com o objetivo de se familiarizar mais com o assunto. No entanto, apesar da seleção inicial, a maioria desses trabalhos não apresenta uma correlação de 100% com o tema central desta pesquisa. Dessa forma, apenas quatro estudos foram considerados altamente relevantes, pois possuem pontos de convergência significativos com o escopo deste trabalho. Os principais aspectos em comum identificados nesses trabalhos, serão analisados e apresentados nos itens 2.1, 2.2, 2.3 e 2.4, destacando suas contribuições para a compreensão do desenvolvimento de um *Data Warehouse,* para análise da produção científica e sua relação com a integração de dados e ferramentas de *Business Intelligence.*

2.2.1 TRABALHO DE MACHADO E MEIRELLES (2005)

O estudo de Machado e Meirelles (2005) analisou a produção científica dos docentes da Universidade Federal da Bahia (UFBA), especificamente na área de Filosofia e Ciências Humanas, no período de 1995 a 1999, identificando os principais produtores de conhecimento, os veículos de divulgação mais utilizados e a distribuição da produtividade entre os departamentos.

O levantamento utilizou como fontes de dados o Catálogo de Produção Científica, Literária e Artística da UFBA (1994-1996) e o Sistema de Produção Docente (SIDoc), permitindo uma análise quantitativa detalhada da produção acadêmica. A pesquisa revelou que, do total de 372 docentes, 63,17% foram responsáveis pela publicação de 2083 trabalhos, incluindo livros, capítulos de livros, artigos em periódicos e comunicações em eventos científicos.

Para ilustrar a distribuição, os autores apresentam a Tabela 04, que evidencia que os professores com titulação de doutorado foram os mais produtivos, responsáveis por 76,48% das publicações. Os docentes com mestrado responderam por 20,02%, enquanto aqueles com apenas graduação ou especialização tiveram um número

significativamente menor de publicações. Essa estatística reforça a correlação entre titulação acadêmica e produtividade científica, demonstrando que docentes inseridos em programas de pós-graduação tendem a produzir mais.

Tabela 3 - Produção por Titulação dos Docentes da Área de Filosofia e Ciências Humanas/UFBA, Período de 1995 a 1999.

Titulação	Produção Absoluta	Produção Relativa	Média no Período
Graduação	54	2.59	10.8
Especialização	19	0.91	3.8
Mestrado	417	20.02	83.4
Doutorado	1593	76.48	318.6
Total	**2083**	**100.00**	

Fonte: Machado e Meirelles (2005).

A pesquisa também destacou a preferência dos docentes pelos seguintes meios de divulgação da produção científica:

- Artigos em periódicos: 33,41% do total das publicações.
- Comunicações em eventos: 18,24%.
- Capítulos de livros: 13,59%.
- Livros completos: 4,85%.

O idioma predominante das publicações foi o português (96,54%), com uma presença discreta de trabalhos em inglês (1,63%), espanhol e francês. Esses dados indicam uma baixa internacionalização da produção científica na área de Filosofia e Ciências Humanas, um fator que pode limitar o alcance e impacto das pesquisas no cenário global.

Outro ponto relevante do estudo foi a análise da distribuição da produção entre os departamentos. Os cinco departamentos mais produtivos da UFBA no período analisado foram:

1. Psicologia (10,56% das publicações).
2. Comunicação (10,23%).

3. Sociologia (9,94%).
4. Educação I (9,70%).
5. Antropologia e Etnologia (9,12%).

Esses departamentos foram responsáveis por quase 50% da produção total da instituição, demonstrando uma concentração de produtividade em determinadas unidades acadêmicas.

Além disso, o estudo categorizou os docentes em quatro grupos de produtividade:

- Grandes produtores: Apenas 3 docentes, responsáveis por 160 publicações (média de 53,33 trabalhos por professor).
- Produtores moderados: 40 docentes, com um total de 929 publicações.
- Transeuntes: 37 docentes, que publicaram entre 9 e 14 trabalhos no período.
- Ocasionais: Representando 65,95% do total de docentes, esses professores publicaram até 8 trabalhos no período de cinco anos.

Entende-se que o estudo de Machado e Meirelles (2005) é altamente relevante para esta pesquisa, pois destaca a importância da análise quantitativa da produção científica para a gestão acadêmica e para o planejamento estratégico das instituições de ensino superior. No contexto do desenvolvimento de um *Data Warehouse* para análise da produção científica, os dados apresentados reforçam a necessidade de ferramentas que possibilitem monitoramento contínuo da produtividade docente, integrando informações de diferentes fontes, como a Plataforma *Lattes* e o Portal da Transparência, para facilitar a tomada de decisões e a visualização estratégica da pesquisa acadêmica.

Ao utilizar soluções de *Business Intelligence* (BI), as universidades podem identificar gargalos na produção científica, mapear áreas estratégicas para investimento e incentivar a internacionalização da pesquisa. Assim, este estudo contribui diretamente para a justificativa do desenvolvimento de um sistema integrado para análise da produção acadêmica, possibilitando a automação e a melhoria dos processos de avaliação e gestão da produtividade docente.

2.2.2 TRABALHO DE ANDRADE ET AL. (2010)

O trabalho de Andrade et al. (2010) teve como objetivo mapear o perfil dos docentes que atuam no Programa de Mestrado em Educação Física da Escola Superior de Educação Física da Universidade Federal de Pelotas (ESEF/UFPel), a partir das informações disponibilizadas nos Currículos *Lattes*. O estudo buscou identificar aspectos relacionados à formação acadêmica, trajetória profissional e inserção dos docentes em linhas e grupos de pesquisa.

Quanto aos procedimentos metodológicos, trata-se de um estudo descritivo, caracterizado como análise documental. A investigação foi realizada com base nos currículos *Lattes* de dezessete docentes, vinculados ao programa de pós-graduação, todos com titulação de doutorado em diferentes áreas do conhecimento. Os dados analisados referem-se à formação inicial e continuada, áreas de atuação, linhas de pesquisa, participação em grupos de pesquisa e produção bibliográfica. Não foi empregado tratamento estatístico avançado, sendo a análise conduzida de forma descritiva e interpretativa a partir das informações curriculares.

Os resultados indicam que o corpo docente apresenta formação heterogênea, tanto no nível de graduação quanto de pós-graduação, com titulações distribuídas entre áreas como Educação Física, Educação, Epidemiologia e Fisiologia. Observou-se também uma distribuição desigual dos docentes entre as linhas de pesquisa do programa, o que reflete escolhas profissionais associadas a interesses acadêmicos e afinidades temáticas.

O estudo aponta ainda que a maioria dos docentes possui forte inserção em grupos de pesquisa cadastrados no Diretório de Grupos de Pesquisa do CNPq, atuando como líderes ou pesquisadores. Essa participação está associada a maior produção científica e ao fortalecimento das redes de colaboração acadêmica, contribuindo para a qualificação do programa de pós-graduação analisado.

Andrade et al. (2010) destacam que o Currículo *Lattes* possibilita a identificação da trajetória acadêmica e profissional dos docentes, permitindo mapear padrões de formação, atuação e produção científica. No entanto, as autoras reconhecem que determinadas atividades docentes não são plenamente contempladas na base curricular, o que limita a compreensão integral da atuação profissional a partir apenas

dos dados registrados.

Dessa forma, o estudo contribui ao demonstrar o potencial do Currículo *Lattes* como fonte documental para análises do perfil docente e da produção acadêmica. Ao mesmo tempo, evidencia a necessidade de abordagens que organizem e sistematizem essas informações de forma estruturada, ampliando as possibilidades de análise e interpretação dos dados curriculares em contextos institucionais e de gestão acadêmica.

2.2.3 TRABALHO DE FERREIRA E SILVA JÚNIOR (2016)

A produção científica é um fator determinante na avaliação de programas de pós-graduação e no reconhecimento institucional dos pesquisadores. No entanto, a pressão para atender às métricas estabelecidas pelas agências de fomento pode impactar diretamente a dinâmica de trabalho dos docentes, especialmente aqueles em início de carreira. O estudo de Ferreira e Silva Júnior (2016) analisou a produção científica de jovens-doutores vinculados à Universidade Federal de São Carlos (UFSCar), considerando aspectos como o ingresso na pós-graduação, participação em projetos de pesquisa e impacto do produtivismo acadêmico.

A pesquisa foi conduzida a partir da análise de professores que possuíam Bolsa Produtividade em Pesquisa (PQ) e menos de dez anos de doutoramento. A metodologia adotada incluiu estudo teórico, levantamento de dados em bases estatísticas oficiais e no Currículo *Lattes*, além da análise empírica do discurso dos docentes. Os resultados evidenciam a existência de um ambiente acadêmico marcado por exigências elevadas de produtividade e internacionalização, refletindo-se na priorização da publicação em periódicos de alto impacto.

Ferreira e Silva Júnior (2016) descrevem o perfil dos jovens-doutores envolvidos, sujeitos da pesquisa, que foi organizada na Tabela 03, que agrupa informações essenciais sobre sua formação e trajetória acadêmica. Para os autores, os jovens-doutores selecionados para esta pesquisa, além de formados a partir de 2004 e estarem com até 10 anos de carreira docente na UFSCar, atuam na pós-graduação e já gozam do status de pesquisadores reconhecidos com bolsa Produtividade em Pesquisa (PQ/CNPq), também chamada de ‘Bolsa PQ’.

A Tabela 03 contém informações sobre os participantes da pesquisa, incluindo a área

de atuação, o ano de doutoramento, a quantidade de pós-doutorado completado, o ano de ingresso como docente na UFSCar, além das categorias e níveis da bolsa Produtividade em Pesquisa (PQ) a que estão vinculados.

Tabela 4 - Identificação dos Jovens-Doutores com bolsa Produtividade em Pesquisa, por códigos de identificação, ano de doutorado, quantidade de pós-doutoramento e ingresso na UFSCar.

Cód. Sujeitos da Pesquisa	**Grande Área**	**Ano Doutorado**	**Num Pós-Doc**	**Pós-Doc 1**	**Pós-Doc 2**	**Concurso UFSCAR**	**Bolsa PQ (LATTES)**
CE1.0509	Ciências Exatas e da Terra	2005	1	2006		2009	PQ2
CE2.0508		2005	4	2007	2008	2008	PQ2
CE3.0404		2004	2	2004	2010	2004	PQ2
CE4.0709		2007	1	2009		2009	PQ2
CE5.0406		2004	2	2004	2006	2006	PQ2

CE6.0408		2004	2	2006	2008	2008	PQ2
CE7.0409		2004	4	2005	2006; 2008	2009	PQ2
CE8.0409		2006	2	2008	2009	2009	PQ2
CE9.0609		2006	2	2006	2009	2009	PQ2
CE10.0509		2005	3	2005	2007	2009	PQ2
CE11.0610		2006	2	2007	2008	2010	PQ2
CE12.0406		2004	2	2006	2013	2006	PQ2
CE13.0710		2007	2	2008	2009	2010	PQ1D
CB14.0809	Ciências Biológicas	2008	1	2009	2009	2009	PQ2
CB15.0608		2006	1	2008		2008	PQ2
ENG16.0410	Engenharias	2004	1	2006		2010	PQ2
ENG17.0609		2006	0			2009	PQ2
ENG18.0609		2006	1	2009		2009	PQ2
ENG19.0404		2004	2	2007	2008	2004	PQ1D
ENG20.0806		2008	0			2006	PQ2
CS21.0404	Ciências da Saúde	2004	1	2008		2004	PQ1C
CS22.0604		2006	1	2014		2004	PQ2
CSA23.0402	Ciências Sociais Aplicadas	2004	0			2002	PQ2
CH24.0405	Ciências Humanas	2004	0			2005	PQ2
CH25.0809		2008	1	2009		2009	PQ2
CH26.0410		2004	0			2010	PQ2
CH27.0606		2006	1	2011		2006	PQ2

Fonte: Ferreira e Silva Júnior (2016).

A análise do cotidiano dos professores revela uma carga de trabalho intensificada, com grande parte do tempo dedicada à submissão de projetos, busca de financiamento e participação em colaborações internacionais. A necessidade de manter um volume expressivo de publicações, aliada à competição por recursos, resulta em um ambiente de alta pressão, o que pode impactar a qualidade de vida dos indivíduos. Ferreira e Silva Júnior (2016), também destacam que, a inserção de

jovens-doutores na pós-graduação tornou-se um ponto central nas políticas de expansão do ensino superior, sendo justificada tanto pela necessidade de qualificação docente quanto pela demanda por aumento da produção científica.

Por fim, o estudo reforça a importância de ferramentas tecnológicas que são capazes de monitorar a produção acadêmica, permitindo uma melhor compreensão das dinâmicas de produtividade e auxiliando gestores na tomada de decisões estratégicas.

2.2.4 TRABALHO DE ESTÁCIO (2017)

O trabalho de Estácio (2017) analisou o Currículo *Lattes* como instrumento de apoio à gestão e à tomada de decisão nas áreas de Ciência, Tecnologia e Inovação no Brasil. A autora caracteriza o Currículo *Lattes* como um sistema de informação responsável por registrar a trajetória acadêmica, científica e profissional dos pesquisadores brasileiros, sendo amplamente utilizado no contexto institucional e governamental.

Quanto aos procedimentos metodológicos, o estudo possui abordagem qualitativa e caráter exploratório, fundamentado em pesquisa bibliográfica e documental. Não é realizada análise estatística nem tratamento quantitativo de dados extraídos diretamente da base do Currículo *Lattes*. A investigação concentra-se na análise histórica e conceitual da Plataforma *Lattes*, com base em documentos institucionais, literatura científica e dados secundários divulgados por órgãos oficiais, como o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

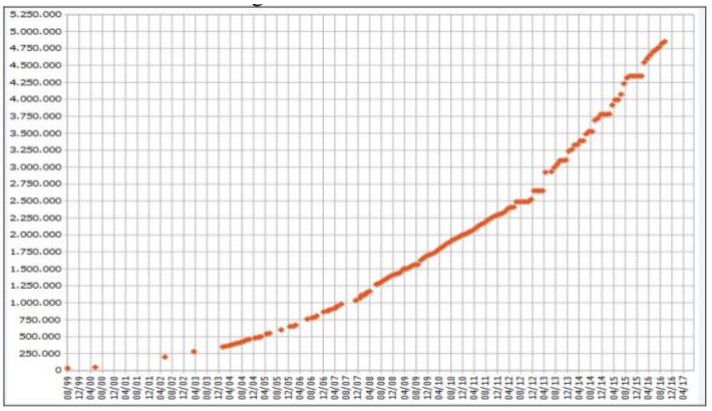
Os resultados apresentados indicam que o Currículo *Lattes* consolidou-se como a principal base nacional de informações sobre pesquisadores, sendo utilizado em processos de avaliação acadêmica, concessão de bolsas, financiamento de projetos de pesquisa e definição de políticas públicas. A autora utiliza dados institucionais previamente publicados para demonstrar o crescimento contínuo da base de currículos ao longo dos anos, evidenciando a expansão e a abrangência da plataforma no cenário científico nacional.

O estudo também aponta que o uso do Currículo Lattes em pesquisas acadêmicas ocorre, predominantemente, de forma descritiva, servindo como fonte para coleta de informações pontuais. Segundo Estácio (2017), são limitadas as iniciativas que exploram essa base de dados a partir de abordagens analíticas estruturadas, capazes

de gerar indicadores consolidados para apoio à gestão científica.

De acordo com matéria publicada no Jornal da Ciência, desde sua criação a Plataforma *Lattes* apresenta crescimento contínuo de sua base de dados, evidenciando sua consolidação no cenário nacional de ciência, tecnologia e inovação (GOULART; TOLENTINO, 2004). Esse crescimento permanece ao longo do tempo, visto que o Currículo *Lattes* passou a concentrar milhões de registros, conforme ilustrado na Figura 1, reforçando a importância da plataforma como um repositório estratégico de informações acadêmicas, científicas e profissionais no contexto da pesquisa científica brasileira.

Figura 1 - Evolução do número de currículos registrados na base do Currículo *Lattes*.

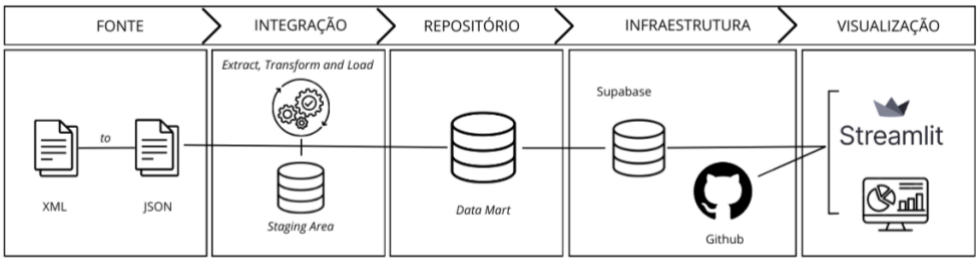


Fonte: Adaptado de (Estácio, 2017).

3. METODOLOGIA

A metodologia adotada neste trabalho, representada na Figura 2, fundamenta-se na arquitetura de *Business Intelligence* proposta por Kimball (2013), a qual orienta a organização de ambientes analíticos baseados em modelagem dimensional. Inicialmente, foram identificadas e analisadas as fontes de dados, compostas por arquivos em formato XML provenientes dos currículos da Plataforma *Lattes*. Em seguida, os dados passaram por um processo de Extração, Transformação e Carga (ETL), no qual foram realizadas atividades de limpeza, padronização e integração, com apoio de uma área intermediária de armazenamento (*Staging Area*), utilizada para preparar os dados antes de sua carga definitiva. Por fim, as informações consolidadas foram armazenadas em um *Data Warehouse*, a partir do qual foram desenvolvidos *dashboards* interativos, possibilitando a visualização dos indicadores e a análise estruturada da produção científica.

Figura 2 - Arquitetura de BI proposta.



Fonte: O autor.

Inicialmente foi realizada uma avaliação detalhada das demandas e metas do projeto, visando identificar os parâmetros e medidas mais importantes para o desenvolvimento da solução de um BI. Uma vez identificada a fonte de dados, foi realizado um estudo que possibilitou a definição de quais informações seriam utilizadas, especificamente quanto aos objetivos da presente proposta. A fonte de dados foi obtida junto à Superintendência de Tecnologia da Informação da Ufes. Foram concedidos arquivos XML dos currículos da Plataforma *Lattes*, os quais foram processados e transformados para o formato JSON, onde foram selecionados apenas os campos de interesse, como exemplo o id *lattes* do pesquisador, suas respectivas áreas e linhas de atuação, bem como artigos e demais produções acadêmicas, entre outros dados relevantes para a análise da produção científica dos docentes.

Para a integração dos dados, utilizou-se uma base intermediária (*Staging Area*).

Segundo Anand e Kumar (2013), a utilização de uma *Staging Area* contribui para a qualidade do processo de ETL, ao permitir a organização e o tratamento dos dados antes de sua carga no repositório final. No contexto deste trabalho, a *staging* area foi utilizada para armazenar os dados extraídos de uma única fonte, disponibilizados em formato JSON, os quais foram previamente transformados e organizados em tabelas relacionais. Essas tabelas foram armazenadas em um banco de dados relacional *PostgreSQL*[4].

Após os dados serem carregados na *staging area*, inicia-se a fase de transformação. Nessa etapa, os dados são sumarizados, integrados e preparados para carga no *Data Warehouse* (DW). O processo de limpeza e padronização foi realizado por meio de duas abordagens complementares. A principal técnica adotada consistiu no uso de instruções SQL do tipo *CASE WHEN* diretamente nas consultas de inserção, permitindo o tratamento de valores nulos ou inconsistentes, a substituição por valores padronizados e a normalização de nomenclaturas institucionais. Adicionalmente, foram empregadas expressões regulares no PostgreSQL para identificação e padronização de variações textuais, considerando diferenças de grafia e acentuação.

Na camada de *staging*, utilizou-se Python[5] para tratamentos básicos de strings, como controle de tamanho dos campos, remoção de valores inválidos e ajustes de formato durante o *parsing* dos arquivos JSON. Após essas etapas, procedeu-se à modelagem dimensional, com a definição de tabelas fato e dimensão, sendo as tabelas fato responsáveis pelo armazenamento das métricas quantitativas do negócio, e as tabelas dimensão destinadas aos atributos descritivos que contextualizam essas informações, e à posterior carga dos dados no DW, viabilizando consultas analíticas eficientes e estruturadas.

Para viabilizar o acesso remoto aos dados e a hospedagem da aplicação de *Business Intelligence*, foi definida uma camada de infraestrutura em nuvem. Os dados armazenados no *Data Warehouse* foram migrados para um banco de dados *PostgreSQL* em ambiente *cloud*, utilizando a plataforma Supabase[6]. O código-fonte da aplicação, foi versionado e disponibilizado no repositório do GitHub[7], onde foi feito

[4] https://www.postgresql.org/
[5] https://www.python.org/
[6] https://supabase.com/
[7] https://github.com/docs

o *deploy* da aplicação no Streamlit[8].

Por fim, a etapa de análise e visualização dos dados foi conduzida por meio da construção de *dashboards* interativos desenvolvidos com a ferramenta Streamlit. As visualizações foram desenvolvidas com o objetivo de responder questões essenciais, abordando a produção acadêmica dos docentes, a distribuição das pesquisas e outros indicadores relevantes.

3.1 FERRAMENTAS UTILIZADAS

3.1.1 Python

Python é uma linguagem de programação interpretada e de alto nível, desenvolvida por Guido van Rossum em 1991 (WIKIPÉDIA, 2026). Neste projeto, a linguagem é utilizada para extrair, transformar e carregar (ETL) os dados dos currículos Lattes. Os *scripts* Python desenvolvidos realizam a leitura dos arquivos XML e transformam cada um deles, por meio da biblioteca *xmltodict*, para o formato JSON, contendo as informações dos pesquisadores. Em seguida, os dados referentes às produções científicas (artigos, livros, capítulos, apresentações de trabalho, projetos de pesquisa, entre outros) são processados.

A linguagem também é utilizada para realizar a limpeza dos dados, principalmente por meio do uso de expressões regulares (*import re*[9]), e para carregar os dados no banco de dados PostgreSQL. Adicionalmente, Python é empregado para popular as tabelas fato do modelo dimensional, realizando as agregações necessárias para a análise.

Entre suas principais características estão a simplicidade e a legibilidade do código, além de possuir uma vasta biblioteca padrão (como *psycopg2* [10]para integração com PostgreSQL e JSON para manipulação de dados) e uma comunidade ativa de desenvolvedores, fatores que a tornam amplamente utilizada em projetos de ETL e ciência de dados (PYTHON SOFTWARE FOUNDATION, 2024).

3.1.2 PostgreSql

Conforme *The PostgreSQL Global Development Group* (2024), o PostgreSQL é um

[8] https://streamlit.io/
[9] https://docs.python.org/pt-br/3/library/re.html
[10] https://pypi.org/project/psycopg2/

sistema gerenciador de banco de dados objeto-relacional de código aberto, que teve sua origem na Universidade da Califórnia em Berkeley. O sistema se destaca por sua robustez, confiabilidade e capacidade de extensão, oferecendo suporte a consultas complexas, chaves estrangeiras e integridade transacional. Estas características, somadas à sua capacidade de processar grandes volumes de dados, o tornam uma escolha apropriada para aplicações de *Data Warehouse*.

3.1.3 Supabase

O Supabase é uma plataforma *open source* que oferece uma camada de serviços em nuvem baseada em PostgreSQL, permitindo a criação, hospedagem e gerenciamento de bancos de dados de forma simplificada. A ferramenta fornece recursos como autenticação, armazenamento e *APIs* automáticas, mantendo total compatibilidade com o ecossistema PostgreSQL.

3.1.4 Streamlit

O Streamlit é uma biblioteca de código aberto da linguagem Python, voltada ao desenvolvimento de aplicações web interativas para ciência de dados e aprendizado de máquina. Ela possibilita o desenvolvimento de aplicações analíticas e *dashboards* a partir de scripts Python. Além disso, o Streamlit facilita a disponibilização e o compartilhamento das aplicações desenvolvidas, sendo amplamente utilizado em projetos de análise e visualização interativa de dados.

3.1.5 Base de dados de origem

A Plataforma *Lattes* é um sistema de informação mantido pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), destinado ao registro e à disponibilização de informações acadêmicas, científicas e profissionais de pesquisadores, docentes e estudantes no Brasil. A plataforma reúne dados sobre formação acadêmica, produção científica, atuação profissional, projetos de pesquisa e participação em grupos de pesquisa, sendo amplamente utilizada por instituições de ensino, agências de fomento e órgãos governamentais para fins de avaliação e gestão científica.

O CNPq disponibiliza às instituições de ensino e pesquisa a possibilidade de extração de dados da Plataforma *Lattes*, tanto da base de Currículos *Lattes* quanto do Diretório

dos Grupos de Pesquisa, por meio de uma ferramenta denominada *Lattes Extrator*. O acesso a essa ferramenta é restrito a instituições previamente cadastradas e requer o atendimento a critérios técnicos e institucionais, como a posse de endereço de IP dedicado e o desenvolvimento de solução tecnológica para integração com o *webservice* disponibilizado.

A Universidade Federal do Espírito Santo (UFES) possui cadastro ativo e acesso ao *Lattes Extrator* de posse da Superintendência de Tecnologia da Informação (STI/UFES). A partir desse acesso, foi disponibilizado um conjunto de dados em formato XML, contendo informações públicas dos currículos da Plataforma *Lattes*, atualmente utilizados pela Superintendência de Projetos e Inovação (SPIN) da UFES. O compartilhamento dos arquivos ocorreu por meio da ferramenta *FileSender*[11], totalizando, em média, 57 mil currículos, em sua grande maioria pertencentes a docentes da Universidade Federal do Espírito Santo (UFES).

Contudo, antes de disponibilizar os dados para utilização na pesquisa, a equipe da Superintendência de Projetos e Inovação (SPIN) da UFES realizou um processo prévio de tratamento, no qual foram removidas informações pessoais sensíveis, garantindo a adequação às normas de proteção de dados. Os arquivos XML resultantes desse processo constituem a base de dados de origem utilizada para o desenvolvimento das etapas de Extração, Transformação e Carga (ETL) descritas nesta metodologia.

3.1.6 Dbeaver

O DBeaver[12] é uma ferramenta utilizada para acessar e administrar diferentes sistemas gerenciadores de bancos de dados, como MySQL, Microsoft SQL Server e PostgreSQL, além de permitir conexões com bancos de dados hospedados em ambientes de nuvem. Trata-se de uma solução de código aberto, desenvolvida em Java, que oferece diferentes versões para atender a variados perfis e necessidades de uso.

Além disso, o DBeaver é amplamente utilizado para o gerenciamento e a manipulação de bancos de dados por meio de uma interface gráfica unificada. No contexto deste

[11] https://filesender.rnp.br/
[12] https://dbeaver.io/

trabalho, a ferramenta é empregada como apoio à execução de consultas SQL e à inspeção da estrutura das tabelas auxiliando no acompanhamento das etapas de carga e transformação dos dados.

4. RESULTADOS OBTIDOS

Com o objetivo de validar a arquitetura de BI desenvolvida, foram implementados cinco *dashboards* analíticos, cada um direcionado a um aspecto específico da análise, permitindo a exploração interativa das informações e a aplicação de filtros conforme a necessidade do usuário. As visões contém filtros que permite a interação do usuário, por meio da seleção e refinamento dos dados.

4.1 LEVANTAMENTO DOS INDICADORES

A partir da análise dos dados extraídos do Currículo *Lattes*, foi possível identificar os conjuntos de informações que melhor representam a atuação acadêmica e científica dos pesquisadores. Com base nessa análise, concluiu-se que os indicadores selecionados são os mais adequados para a construção do *Business Intelligence* proposto, possibilitando a geração de análises consistentes, a identificação de padrões e tendências, bem como a exploração da produção científica sob diferentes perspectivas. No presente estudo, os indicadores foram organizados nos seguintes grupos:

- Distribuição por grandes áreas e áreas específicas.
- Pesquisadores com atuação em múltiplas áreas e em área única.
- Distribuição da produção por grande área.
- Média de pesquisadores por linha de pesquisa.
- Evolução temporal e tendências da produção científica.
- Distribuição das produções por tipo.
- Produtividade média dos pesquisadores.
- Quantidade de pesquisadores ativos por ano.
- Crescimento da base de pesquisadores ao longo do tempo.
- Produções e apresentações realizadas na UFES.
- Percentual de internacionalização da produção científica.

4.2 ESTRUTURA DO PROJETO

O projeto foi desenvolvido com o auxílio do editor *Visual Studio Code*[13] (Visual Studio Code, 2026), utilizado para a implementação e organização do código-fonte. A estrutura de diretórios, apresentada na Figura 3, evidencia a disposição dos módulos responsáveis pelas etapas de processamento, transformação e análise dos dados, bem como dos demais arquivos que integram a aplicação. Todo o código desenvolvido está disponibilizado em um repositório no *GitHub*[14], possibilitando o acesso e o versionamento do projeto.

De acordo com Gamma et al. (1994), a adoção de padrões contribui para uma melhor comunicação entre os componentes de um sistema, facilitando sua reutilização, compreensão e manutenção ao longo do tempo.

Figura 3 - Estrutura do projeto no Visual Studio Code.

Fonte: O autor.

O detalhamento de cada uma das pastas que compõem essa estrutura, ilustrada na Figura 3, é apresentado nas seções a seguir, descrevendo suas respectivas funções no contexto do processo de ETL e da implementação do *Business Intelligence* proposto.

4.2.1 Pasta db

Diretório responsável pela camada de acesso ao banco de dados PostgreSQL. Nele

[13] https://code.visualstudio.com/
[14] https://github.com

estão concentradas as rotinas de conexão, bem como os parâmetros necessários para autenticação e comunicação com o banco. Além disso, contém funções auxiliares utilizadas para a execução de consultas SQL e para o controle de transações durante o processamento dos dados.

4.2.2 Pasta lattes_tcc

Diretório destinado ao armazenamento dos currículos extraídos da Plataforma *Lattes*, mantidos em formato JSON. Esse diretório reúne um total de 57.062 arquivos, que constituem a base de dados bruta utilizada como fonte primária para as etapas de extração, transformação e carga dos dados no projeto.

4.2.3 Pasta stage

Contém os scripts desenvolvidos em Python responsáveis pela etapa de ETL. Esses scripts realizam a leitura e o processamento dos arquivos JSON provenientes do Currículo *Lattes*, efetuando a transformação dos dados e o carregamento em estruturas intermediárias no banco de dados. O diretório é composto por módulos específicos para diferentes tipos de informações e produções científicas, incluindo artigos publicados, livros e capítulos, trabalhos em eventos, apresentações, textos em jornais e revistas, outras produções técnicas, projetos de pesquisa, áreas de atuação, linhas de pesquisa e dados cadastrais dos pesquisadores.

4.2.4 Pasta sql_criar_tabelas

Diretório que reúne os scripts SQL responsáveis pela criação da estrutura do *Data Warehouse*. Nele estão definidas as tabelas dimensionais, as tabelas fato e as tabelas utilizadas na etapa de *staging*, organizadas em arquivos distintos conforme sua finalidade dentro do modelo dimensional proposto.

4.2.5 Pasta populando_tabelas

Abriga os scripts em Python encarregados de realizar a carga dos dados no *Data Warehouse*. Esses scripts efetuam a transferência dos dados tratados a partir do banco intermediário para as tabelas dimensionais e tabelas fato, além de contemplarem rotinas de validação, verificação de inconsistências e criação de índices, visando garantir a integridade dos dados e a otimização das consultas analíticas.

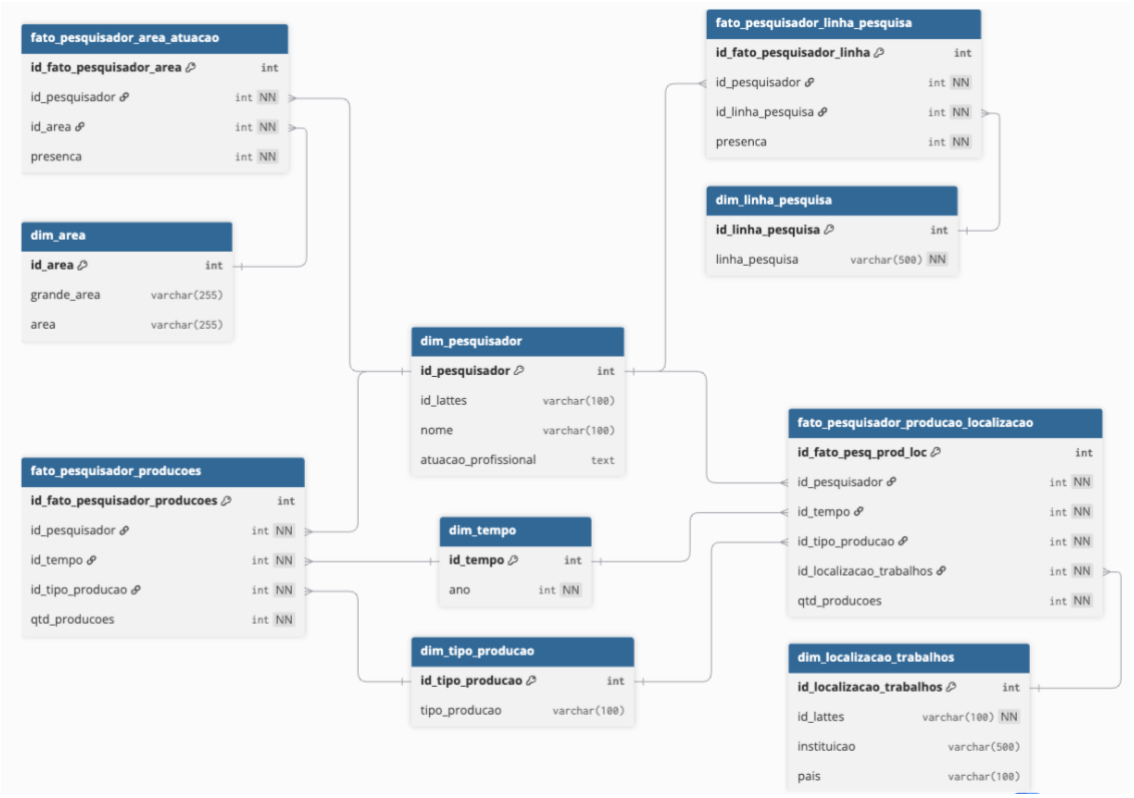
4.2.6 Pasta streamlit

Diretório dedicado à aplicação web de *Business Intelligence* desenvolvida com a ferramenta Streamlit. Nele estão implementadas as interfaces de visualização e análise dos dados, incluindo cinco dashboards temáticos organizados por área de análise. O diretório também contempla configurações visuais alinhadas à identidade institucional da UFES, componentes reutilizáveis para padronização da interface, arquivos de estilo e documentação técnica relacionada ao processo de implantação da aplicação.

4.3 MODELO DIMENSIONAL

A Figura 4 ilustra o modelo dimensional proposto neste trabalho, permitindo a compreensão da organização e do relacionamento entre os elementos que compõem o *Data Warehouse* desenvolvido. Outros detalhes sobre as estruturas definidas podem ser vistos no dicionário de dados disponível no Apêndice A.

Figura 4 - Modelo Dimensional.



Fonte: O autor.

No contexto deste estudo, foram definidas seis tabelas dimensionais, as quais são

apresentadas a seguir:

4.3.1 dim_pesquisador

Armazena informações sobre os pesquisadores cadastrados na Plataforma Lattes. Contém o identificador único do *Lattes*, nome completo e atuação profissional de cada pesquisador.

4.3.2 dim_area

Representa as áreas de conhecimento em que os pesquisadores atuam, seguindo a classificação do CNPq. Inclui a grande área.

4.3.3 dim_linha_pesquisa

Contém as linhas de pesquisa declaradas pelos pesquisadores em seus currículos *Lattes*.

4.3.4 dim_tempo

Dimensão temporal que armazena os anos de publicação das produções científicas.

4.3.5 dim_tipo_producao

Classifica os diferentes tipos de produção científica, contemplando artigos publicados, livros, capítulos de livros, trabalhos em eventos, apresentações de trabalhos, projetos de pesquisa, textos publicados em jornais e outras produções técnicas.

4.3.6 dim_localizacao_trabalhos

Armazena dados geográficos associados às produções científicas, contemplando a instituição e o país nos quais os trabalhos foram realizados ou publicados.

Adicionalmente, foram criadas quatro tabelas fato, uma vez que a utilização de mais de uma tabela desse tipo mostrou-se essencial para atender às diferentes necessidades analíticas do projeto, permitindo a separação de métricas com granularidades distintas e garantindo maior organização, desempenho e flexibilidade ao modelo dimensional.

4.3.7 fato_pesquisador_producoes

Armazena a quantidade de produções científicas associadas a cada pesquisador,

organizadas por tipo de produção e período temporal.

4.3.8 fato_pesquisador_area_atuacao

Registra a relação entre os pesquisadores e suas áreas de atuação, indicando a presença de cada pesquisador em determinadas áreas do conhecimento.

4.3.9 fato_pesquisador_linha_pesquisa

Representa o vínculo entre os pesquisadores e suas respectivas linhas de pesquisa, evidenciando a participação em diferentes temas de investigação.

4.3.10 fato_pesquisador_producao_localizacao

Integra informações sobre a produção científica registrando o período e a instituição de publicação.

O modelo dimensional foi estruturado a partir de quatro esquemas estrela, cada um associado a uma tabela fato específica e às dimensões necessárias para sua análise. A tabela fato_pesquisador_producoes relaciona-se com as dimensões pesquisador, tempo e tipo de produção, sendo utilizada para análises de produtividade científica ao longo do tempo. A tabela fato_pesquisador_area_atuacao compartilha a dimensão pesquisador e associa-se à dimensão área de atuação, permitindo a análise da distribuição dos pesquisadores por área do conhecimento. A tabela fato_pesquisador_linha_pesquisa também compartilha a dimensão pesquisador, relacionando-se à dimensão linha de pesquisa, possibilitando a análise dos temas de investigação associados aos pesquisadores. Por fim, a tabela fato_pesquisador_producao_localizacao integra as dimensões pesquisador, tempo, tipo de produção e a dim_localização_trabalhos, sendo utilizada para análises que envolvem a distribuição geográfica em instituições das produções científicas.

A Figura 5 (a) apresenta a visão geral da arquitetura adotada, evidenciando a divisão entre os esquemas destinados ao armazenamento dos dados brutos (STG) e ao armazenamento das estruturas analíticas do *Data Warehouse* (DW). A arquitetura do banco de dados foi organizada a partir da separação lógica dos ambientes por meio da utilização de esquemas distintos. Essa estratégia permitiu isolar os dados em estágio inicial de processamento daqueles já estruturados para análise, garantindo maior organização e controle ao longo do processo de ETL. Nas Figuras 5 (b) e 5 (c)

é possível observar, respectivamente, a disposição das tabelas pertencentes à *staging area*, que reúne os dados extraídos e ainda não consolidados, e das tabelas do esquema DW, que contempla as tabelas dimensionais e as tabelas fato responsáveis por suportar as análises propostas neste estudo.

Figura 5 - Estrutura do banco de dados visualizada no DBeaver, destacando os esquemas DW e STG.

Fonte: O autor.

Inicialmente, o banco de dados PostgreSQL foi desenvolvido e testado localmente. Para realizar o acesso remoto pelo Streamlit Cloud, tornou-se necessária a migração do banco para a nuvem, sendo utilizada a plataforma Supabase, que oferece hospedagem gratuita para bancos PostgreSQL. O processo de migração envolveu a criação de um projeto no Supabase, no qual foi disponibilizado um banco de dados PostgreSQL padrão. Em seguida, foram executados os scripts SQL responsáveis pela criação das tabelas dimensão e fato do modelo dimensional, assim como os scripts de carga (ETL) utilizados para a população do *Data Warehouse* com os dados tratados. Além disso, foi realizada a configuração de acesso seguro ao banco de dados, por meio de conexão SSL, garantindo a integridade e a segurança das informações durante o acesso remoto. Para assegurar a compatibilidade com ambientes que operam exclusivamente em IPv4, como o Streamlit Cloud, optou-se pela utilização do mecanismo de *Session Pooler* disponibilizado pelo *Supabase*, que realiza o gerenciamento eficiente das conexões em aplicações hospedadas em nuvem.

4.4 GRÁFICOS

Após a conclusão do processo de carga e organização dos dados no *Data Warehouse*,

deu-se início à etapa de desenvolvimento das visualizações analíticas por meio da ferramenta *Streamlit*. Essa ferramenta possibilita a criação de aplicações web interativas conectadas diretamente ao banco de dados, permitindo a exploração dos dados a partir da definição de filtros e da integração entre as tabelas dimensionais e as tabelas fato do modelo proposto.

Na sequência, são apresentados os *dashboards* desenvolvidos neste estudo, selecionados com base na relevância das análises realizadas. A aplicação completa, contendo todos os *dashboards* implementados, encontra-se disponível para acesso por meio deste *link*[15]*.*

- *Dashboard* 1 – Áreas de Atuação: apresenta a distribuição dos pesquisadores e de suas produções de acordo com as áreas de atuação.
- *Dashboard* 2 – Linhas de Pesquisa: permite analisar as linhas de pesquisa associadas aos pesquisadores.
- *Dashboard* 3 – Evolução Temporal: evidencia a evolução da produção científica ao longo do tempo.
- *Dashboard* 4 – Produtividade e Rankings: apresenta indicadores de produtividade e rankings de pesquisadores.
- *Dashboard* 5 – Localização Geográfica: demonstra a distribuição geográfica dos pesquisadores e de suas produções.

A Figura 6 apresenta a distribuição dos pesquisadores por grande área do conhecimento. O gráfico evidencia a quantidade de pesquisadores associados a cada grande área, permitindo visualizar a concentração da atuação acadêmica nos diferentes campos científicos. Além disso, o painel apresenta indicadores gerais, como o total de pesquisadores analisados e o número total de grandes áreas consideradas na base. Observa-se que, as áreas Ciências da Saúde, Ciências Sociais Aplicadas e Ciências Humanas se destacam com mais de 7% do total de pesquisadores. Esse resultado indica uma predominância dessas áreas no cenário institucional, sugerindo

[15] https://dw-streamlit.app/

maior densidade acadêmica e potencial de produção científica nesses campos.

Em contrapartida, áreas como Ciências Biológicas e Ciências Agrárias apresentam menor representatividade relativa, o que pode refletir diferenças estruturais, históricas ou estratégicas na distribuição de recursos e na consolidação dos programas de pesquisa.

A visualização permite não apenas identificar padrões de concentração, mas também subsidiar análises estratégicas relacionadas à alocação de investimentos, planejamento institucional e definição de políticas acadêmicas.

Figura 6 - Análise da distribuição de Pesquisadores por Grande Área.

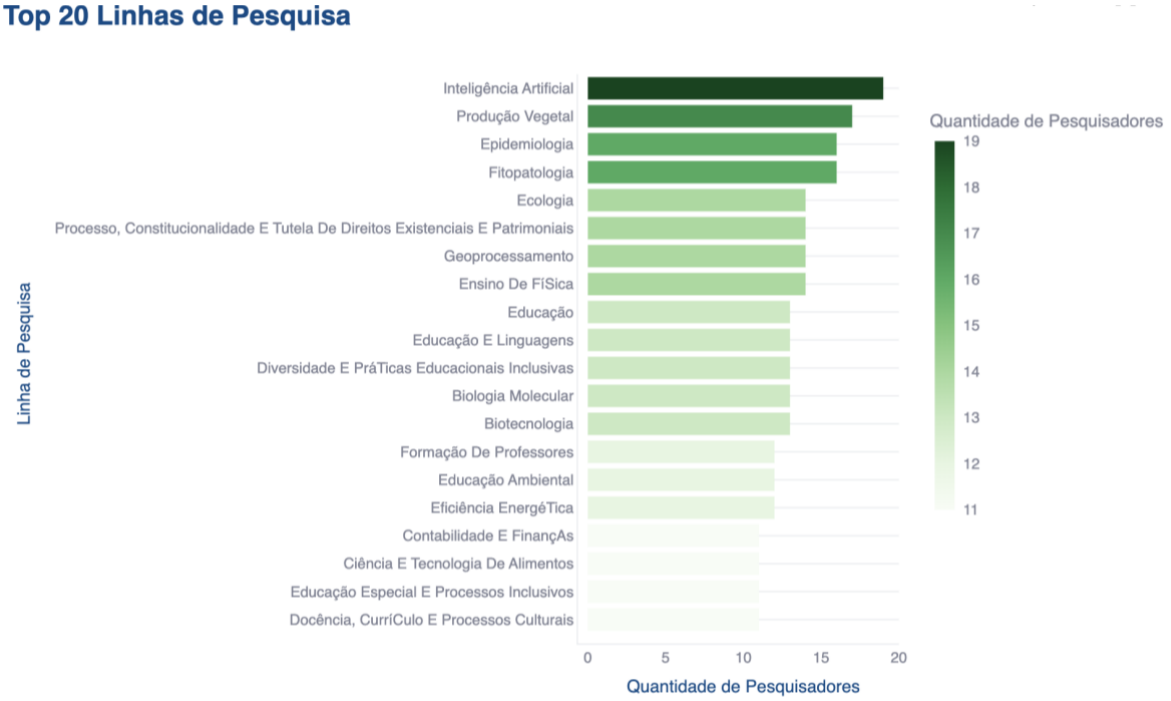
Fonte: O autor.

O gráfico apresentado na Figura 7, apresenta o *ranking* das 20 linhas de pesquisa com maior número de pesquisadores. Esse gráfico permite identificar os temas de pesquisa que concentram maior participação de pesquisadores, oferecendo uma visão comparativa entre as linhas mais recorrentes na amostra analisada. A visualização auxilia na compreensão da distribuição dos pesquisadores entre os diferentes eixos temáticos de investigação científica.

Observa-se que a linha de Inteligência Artificial ocupa a primeira posição no ranking,

evidenciando a crescente relevância dessa área no cenário acadêmico. Esse destaque dialoga com a consolidação global da área nos últimos anos, especialmente após avanços significativos impulsionados por estudos como *Attention Is All You Need* (VASWANI et al., 2017*)*, que marcou uma mudança paradigmática no campo da Inteligência Artificial, especialmente no desenvolvimento de modelos de linguagem de grande porte e em aplicações avançadas de processamento de linguagem natural. Além disso, linhas como Produção Vegetal, Epidemiologia e Fitopatologia também se destacam, indicando a forte presença de pesquisas voltadas às áreas da saúde e do setor agroambiental.

Figura 7 - Ranking das 20 linhas de pesquisa com maior número de pesquisadores.



Fonte: O autor.

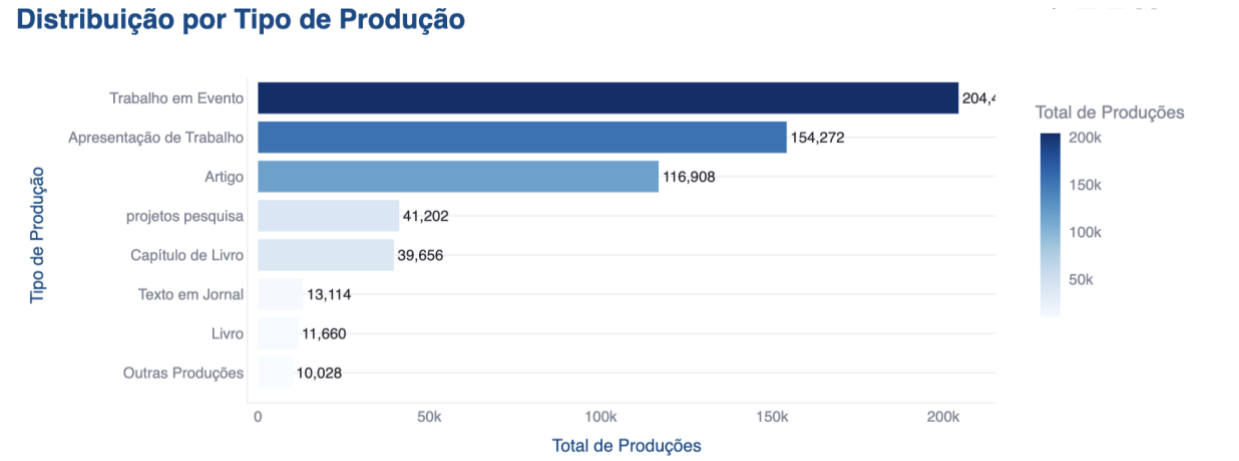
A Figura 8 exibe a distribuição das produções científicas por tipo de produção. O gráfico apresenta o quantitativo total de produções categorizadas em diferentes tipos, como artigos, livros, capítulos de livros, trabalhos em eventos e outras produções. Essa visualização possibilita observar quais tipos de produção são mais frequentes na base de dados analisada.

Observa-se que trabalhos em eventos representam a categoria mais expressiva, ultrapassando 200 mil registros, seguidos por apresentações de trabalho e artigos científicos. Esse resultado evidencia uma forte cultura de divulgação científica em

eventos acadêmicos, indicando intensa participação da comunidade em congressos, simpósios e seminários.

Por outro lado, produções como livros, textos em jornal e outras categorias apresentam quantitativos inferiores, o que pode refletir diferenças no tempo de maturação dessas produções ou nas dinâmicas próprias de cada área do conhecimento.

Figura 8 - Distribuição por tipo de produção.



Fonte: O autor.

A Figura 9 apresenta a evolução anual da quantidade de pesquisadores ativos ao longo do período analisado (1964 à 2025). O gráfico evidencia a variação no número de pesquisadores com produção registrada ao longo dos anos, permitindo a observação de períodos de crescimento ou redução da atividade científica. O painel também apresenta indicadores resumidos, como o período total analisado, a média anual de pesquisadores e o ano com maior número de pesquisadores ativos. Observa-se uma queda acelerada a partir do ano de 2020.

Nota-se um crescimento expressivo e contínuo no número de pesquisadores ativos a partir dos anos 2000, com aceleração mais intensa na década de 2010, culminando no pico registrado em 2019, ano que apresentou o maior quantitativo da série histórica. Esse movimento sugere um período de expansão da atividade científica institucional, possivelmente associado à ampliação de programas de pós-graduação e incentivos à pesquisa. A queda observada após 2020 pode estar relacionada a fatores conjunturais, como restrições orçamentárias e impactos da pandemia, indicando uma inflexão relevante na tendência de crescimento previamente consolidada.

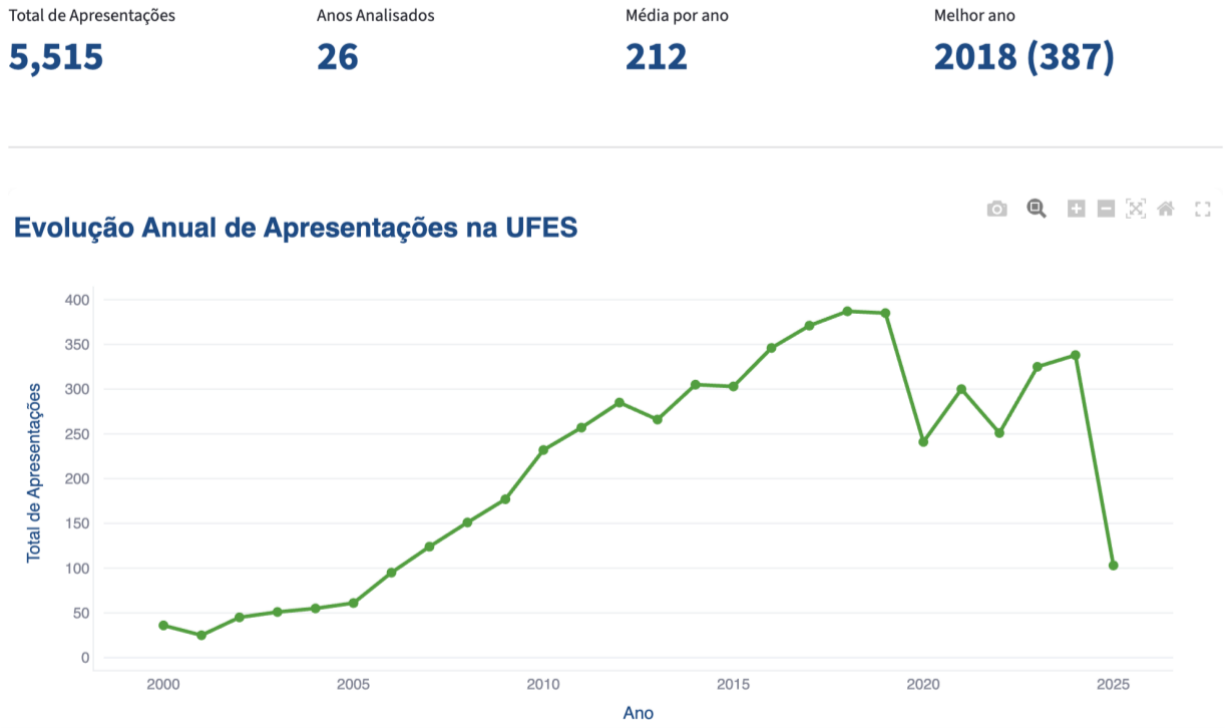
Figura 9 - Evolução anual do número de pesquisadores ativos.

Fonte: O autor.

Por fim, o gráfico representado pela Figura 10, apresenta a evolução anual do número de apresentações de trabalhos realizadas na UFES. O gráfico permite visualizar o comportamento desse tipo de produção ao longo do tempo, destacando variações anuais e identificando os períodos com maior concentração de apresentações registradas na base de dados. Observa-se um crescimento consistente ao longo dos anos, com destaque para 2018, que registrou o maior quantitativo de apresentações.

Nota-se uma queda significativa no período pós-pandemia, refletindo o impacto das restrições a eventos presenciais e atividades acadêmicas. Nos anos subsequentes, verifica-se uma retomada gradual das apresentações, indicando um movimento de recuperação das atividades científicas. O quantitativo reduzido observado em 2025 pode ser explicado pelo fato de os dados terem sido extraídos no mês de setembro, não contemplando o ano completo, o que tende a impactar a representatividade dos registros mais recentes.

Figura 10 - Quantidade de apresentações de trabalho que ocorreram na UFES.



Fonte: O autor.

De forma geral, os gráficos apresentados têm caráter descritivo, permitindo a visualização e a exploração dos dados disponíveis no *Data Warehouse*. Destaca-se que as análises refletem as informações efetivamente preenchidas nos currículos *Lattes*, estando sujeitas a limitações decorrentes de dados ausentes ou incompletos.

5. CONCLUSÕES

O presente trabalho teve como objetivo desenvolver uma solução de *Business Intelligence* aplicada à análise da produção científica de pesquisadores, a partir de dados extraídos do Currículo *Lattes*. Para isso, foi realizada a construção de um *Data Warehouse* baseado em modelagem dimensional, aliado à implementação de rotinas de ETL e ao desenvolvimento de *dashboards* interativos por meio da ferramenta Streamlit.

A organização dos dados em esquemas distintos de *staging* e *Data Warehouse* permitiu estruturar adequadamente as informações, garantindo maior consistência e clareza ao processo analítico. A modelagem dimensional, composta por tabelas de dimensão e múltiplas tabelas fato, mostrou-se adequada para representar diferentes perspectivas da produção científica, como a atuação dos pesquisadores em grandes áreas do conhecimento, linhas de pesquisa, tipos de produção, evolução temporal e localização das produções.

A partir das visualizações desenvolvidas, foi possível observar que a atuação dos pesquisadores encontra-se distribuída entre diversas grandes áreas do conhecimento, com destaque para áreas que concentram maior número de pesquisadores, evidenciando a diversidade da produção científica analisada. Os *dashboards* também permitiram identificar as linhas de pesquisa com maior concentração de pesquisadores, bem como a predominância de determinados tipos de produção científica, como artigos, trabalhos em eventos e apresentações, refletindo diferentes formas de disseminação do conhecimento acadêmico.

Destaca-se, entretanto, que os resultados apresentados devem ser interpretados considerando as limitações da base de dados utilizada. As informações analisadas dependem do preenchimento voluntário dos currículos pelos próprios pesquisadores na Plataforma *Lattes*, o que resulta na presença de campos nulos ou incompletos. Ademais, a análise foi realizada a partir de um subconjunto dos currículos disponíveis, o que pode impactar a abrangência e a representatividade dos indicadores apresentados.

Ainda assim, o estudo evidencia o potencial do uso de técnicas de *Business Intelligence* para a análise da produção científica, demonstrando como a integração

entre *Data Warehouse* e ferramentas de visualização pode apoiar a exploração e a compreensão de grandes volumes de dados acadêmicos. A solução desenvolvida permite a análise dos dados sob diferentes perspectivas, contribuindo para a identificação de padrões, tendências e distribuições relevantes.

Como trabalhos futuros, sugere-se a ampliação da base de dados analisada, o aprimoramento das rotinas de tratamento de dados ausentes e a inclusão de novos indicadores, de modo a enriquecer as análises e aumentar a robustez da solução proposta. Destaca-se ainda diferentes possibilidades de continuidade, seja por meio da incorporação de novas fontes de dados, do aprofundamento analítico ou da evolução da infraestrutura tecnológica adotada. Em síntese, o trabalho atingiu os objetivos estabelecidos, apresentando uma arquitetura de BI funcional, organizada e alinhada às boas práticas da área de dados, constituindo uma base sólida para análises futuras sobre a produção científica no contexto da Universidade Federal do Espírito Santo.

A solução proposta tem o potencial de impactar positivamente as avaliações institucionais, a identificação de áreas de destaque ou melhoria e a visibilidade da pesquisa no cenário nacional e internacional por parte das instituições de ensino superior. Além de ser uma importante fonte de dados, integrada e de fácil acesso, para pesquisas futuras.

6. REFERÊNCIAS

ANAND, N.; KUMAR, M. **An overview on data quality issues at data staging ETL**. In: *International Conference on Advances in Signal Processing and Communication*. [S.l.]:, 2013.

ESTÁCIO, Letícia Silvana dos Santos. A importância do currículo Lattes como ferramenta que representa a ciência, tecnologia e inovação no país. Revista ACB: Biblioteconomia em Santa Catarina, v. 22, n. esp. 2, p. 300-311, 2017. Disponível em: https://dialnet.unirioja.es/servlet/articulo?codigo=6484974. Acesso em: 9 fev. 2026.

FERREIRA, Luciana Rodrigues; SILVA JÚNIOR, João dos Reis. A produção científica do professor-pesquisador jovem-doutor. Trabalho Necessário, Niterói, ano 14, n. 23, p. 118-143, 2016. Disponível em: https://periodicos.uff.br/trabalhonecessario/article/view/9604/6725. Acesso em: 9 fev. 2026.

ANDRADE, Danièlle Müller de et al.O perfil dos docentes da ESEF/UFPEL através do Currículo Lattes. Revista Didática Sistêmica, [S. l.], v. 11, p. 165–177, 2011. DOI: 10.14295/rds.v11i0.1752. Disponível em: https://periodicos.furg.br/redsis/article/view/1752. Acesso em: 22 jan. 2026.

GAMMA, Erich et al. **Design patterns: elements of reusable object-oriented software**. 1. ed. Boston: Addison-Wesley Professional, 1994. ISBN 0201633612.

GEMELLI, Catia Eli; CLOSS, Lisiane Quadrado. Trabalho docente no ensino superior: análise da produção científica publicada no Brasil (2010-2019). Educação & Sociedade, Campinas, v. 43, e246522, 2022. Disponível em: https://doi.org/10.1590/ES.246522. Acesso em: 9 fev. 2026.

PRODUÇÃO científica brasileira cai pelo segundo ano consecutivo. Sociedade Brasileira para o Progresso da Ciência (SBPC), 30 jul. 2024. Disponível em: https://portal.sbpcnet.org.br/noticias/producao-cientifica-brasileira-cai-pelo-segundo-ano-consecutivo/. Acesso em: 9 fev. 2026.

KIMBALL, Ralph; ROSS, Margy. *The data warehouse toolkit: the definitive guide to dimensional modeling*. 3. ed. Indianapolis: Wiley, 2013. Disponível em:

https://www.oreilly.com/library/view/the-data-warehouse/9781118530801/. Acesso em: 9 fev. 2026.

MACHADO, Raymundo das Neves; MEIRELLES, Rodrigo França. Produção científica dos docentes da Universidade Federal da Bahia da área de Filosofia e Ciências Humanas no período de 1995-1999. *Revista Brasileira de Política Internacional*, Brasília, v. 42, n. 2, p. 81-92, 1999. Disponível em: https://www.scielo.br/j/tinf/a/9FpBN96V9wVJSCVLPDmhCnm/?lang=pt. Acesso em: 22 jan. 2025.

REIS, Nadson Santana; CASTRO, Pedro Alves; CARDOSO, Berta Leni Costa; NUNES, Claudio Pinto. A produção de conhecimento sobre trabalho docente no Brasil: uma revisão da literatura especializada no assunto. Entreideias: Revista de Educação, Salvador, v. 9, n. 2, p. 87-112, maio/ago. 2020. Disponível em: https://periodicos.ufba.br/index.php/entreideias/article/view/33242/21867. Acesso em: 09 fev. 2025.

UNIVERSIDADE FEDERAL DO PARÁ (UFPA). Universidades públicas realizam mais de 95% da ciência no Brasil. 18 abr. 2019. Disponível em: https://ufpa.br/universidades-publicas-realizam-mais-de-95-da-ciencia-no-brasil/. Acesso em: 8 mar. 2025.

VIEIRA, Douglas. API: entenda a interface de programação de aplicação. *Blog HostGator*, 2023. Disponível em: https://www.hostgator.com.br/blog/api-entenda-a-interface-de-programacao-de-aplica cao/. Acesso em: 16 jan. 2026.

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR (CAPES). Brasil produz 6,3 mil estudos sobre inteligência artificial. 2024. Disponível em: https://www.gov.br/capes/pt-br/assuntos/noticias/brasil-produz-6-3-mil-estudos-sobre-inteligencia-artificial. Acesso em: 31 jan. 2026.

VAN ROSSUM, Guido. Wikipédia, a enciclopédia livre. Wikipédia: a enciclopédia livre, 2026. Disponível em: https://pt.wikipedia.org/wiki/Guido_van_Rossum. Acesso em: 01 fev. 2026.

NETTO, André. Produção científica brasileira tem reconhecimento mundial. Jornal da USP, 5 jul. 2019. Disponível em: https://jornal.usp.br/?p=256478. Acesso em: 25 fev.

2026.

ANDRADE, Sergio Fred. *Business Intelligence: data warehousing* na prática. [S. l.: s. n.], 2017. 50 p. Disponível em: https://programador1.com/assets/files/e-book-bi-datawarehousing-na-pratica-v8_b2.pdf. Acesso em: 25 fev. 2026.

APÊNDICE A - DICIONÁRIO DE DADOS

Schema	Nome_tabela	Coluna	Tipo_dado	Descrição
stg	pesquisador	id_lattes	varchar(100)	Identificador único do pesquisador no currículo Lattes
stg	pesquisador	nome	varchar(100)	Nome completo do pesquisador
stg	pesquisador	atuacao_profissional	text	Instituição de atuação profissional do pesquisador
stg	linha_pesquisador	id	serial(4)	Identificador único da linha de pesquisa
stg	linha_pesquisador	id_lattes	varchar(100)	Identificador do pesquisador no Lattes
stg	linha_pesquisador	linha_pesquisa	varchar(500)	Nome ou descrição da linha de pesquisa do pesquisador
stg	areas_atuacao	id	serial(4)	Identificador único da área de atuação
stg	areas_atuacao	id_lattes	varchar(100)	Identificador do pesquisador no Lattes
stg	areas_atuacao	nome_grande_area	varchar(255)	Nome da grande área do conhecimento
stg	areas_atuacao	nome_area	varchar(255)	Nome da área específica
stg	areas_atuacao	nome_sub_area	varchar(255)	Nome da subárea de conhecimento
stg	areas_atuacao	nome_especialidade	varchar(255)	Nome da especialidade dentro da subárea
stg	projetos_pesquisa	id	serial(4)	Identificador único do projeto
stg	projetos_pesquisa	id_lattes	varchar(100)	Identificador do pesquisador no Lattes
stg	projetos_pesquisa	ano_inicio	varchar(10)	Ano de início do projeto de pesquisa
stg	projetos_pesquisa	ano_fim	varchar(10)	Ano de conclusão do projeto de pesquisa
stg	projetos_pesquisa	nome_projeto	varchar(500)	Título ou nome do projeto de pesquisa
stg	projetos_pesquisa	descrição_projeto	text	Descrição detalhada do projeto de pesquisa
stg	projetos_pesquisa	situacao	varchar(100)	Situação atual do projeto (em andamento/concluído)

stg	projetos_pesquisa	natureza	varchar(100)	Natureza do projeto (pesquisa/extensão/desenvolvimento)
stg	artigos	id	serial(4)	Identificador único do artigo
stg	artigos	id_lattes	varchar(100)	Identificador do pesquisador no Lattes
stg	artigos	ano	varchar(10)	Ano de publicação do artigo
stg	artigos	titulo	varchar(1000)	Título completo do artigo
stg	artigos	doi	varchar(255)	Digital Object Identifier do artigo
stg	artigos	idioma	varchar(50)	Idioma de publicação do artigo
stg	artigos	natureza	varchar(50)	Natureza da publicação (completo/resumo)
stg	artigos	meio_divulgacao	varchar(50)	Meio de divulgação (impresso/digital)
stg	artigos	titulo_periodico	varchar(500)	Nome da revista ou periódico
stg	artigos	volume	varchar(20)	Volume da revista em que foi publicado
stg	artigos	pagina_inicial	varchar(10)	Número da página inicial do artigo
stg	artigos	pagina_final	varchar(10)	Número da página final do artigo
stg	artigos	issn	varchar(20)	International Standard Serial Number da revista
stg	artigos	local_publicacao	varchar(255)	Local (cidade/país) de publicação
stg	artigos	autores	text	Lista de autores do artigo
stg	livros	id	serial(4)	Identificador único do livro
stg	livros	id_lattes	varchar(100)	dentificador do pesquisador no Lattes
stg	livros	ano	varchar(10)	Ano de publicação do livro
stg	livros	titulo	varchar(1000)	Título completo do livro
stg	livros	numero_edicao	varchar(50)	Número da edição do livro
stg	livros	cidade_editora	varchar(255)	Cidade da editora
stg	livros	nome_editora	varchar(255)	Nome da editora responsável pela publicação

stg	livros	numero_volumes	varchar(50)	Quantidade de volumes do livro
stg	livros	numero_paginas	varchar(50)	Número total de páginas do livro
stg	livros	autores	text	Lista de autores do livro
stg	capitulos_livros	id	serial(4)	Identificador único do capítulo
stg	capitulos_livros	id_lattes	varchar(100)	Identificador do pesquisador no Lattes
stg	capitulos_livros	ano	varchar(10)	Ano de publicação do capítulo
stg	capitulos_livros	titulo_capitulo	varchar(1000)	Título do capítulo escrito
stg	capitulos_livros	titulo_livro	varchar(1000)	Título do livro que contém o capítulo
stg	capitulos_livros	doi	varchar(255)	Digital Object Identifier do capítulo
stg	capitulos_livros	idioma	varchar(50)	Idioma de publicação do capítulo
stg	capitulos_livros	meio_divulgacao	varchar(50)	Meio de divulgação (impresso/digital)
stg	capitulos_livros	numero_edicao	varchar(50)	Número da edição do livro
stg	capitulos_livros	cidade_editora	varchar(255)	Cidade da editora
stg	capitulos_livros	nome_editora	varchar(255)	Nome da editora responsável pela publicação
stg	capitulos_livros	isbn	varchar(50)	International Standard Book Number do livro
stg	capitulos_livros	pagina_inicial	varchar(10)	Número da página inicial do capítulo
stg	capitulos_livros	organizadores	varchar(1000)	Organizadores do livro
stg	capitulos_livros	autores	text	Lista de autores do capítulo
stg	textos_jornais	id	serial(4)	dentificador único do texto
stg	textos_jornais	id_lattes	varchar(100)	Identificador do pesquisador no Lattes
stg	textos_jornais	ano	varchar(10)	Ano de publicação do texto
stg	textos_jornais	titulo	varchar(1000)	Título do texto publicado
stg	textos_jornais	titulo_jornal	varchar(500)	Nome do jornal ou revista

stg	textos_jornais	doi	varchar(255)	Digital Object Identifier do texto
stg	textos_jornais	idioma	varchar(50)	Idioma de publicação
stg	textos_jornais	natureza	varchar(100)	Natureza do texto (artigo de opinião/reportagem/etc)
stg	textos_jornais	meio_divulgacao	varchar(50)	Meio de divulgação (impresso/digital)
stg	textos_jornais	data_publicacao	varchar(20)	Data específica de publicação
stg	textos_jornais	local_publicacao	varchar(255)	Local (cidade/país) de publicação
stg	textos_jornais	pagina_inicial	varchar(10)	Número da página inicial
stg	textos_jornais	pagina_final	varchar(10)	Número da página final
stg	textos_jornais	volume	varchar(20)	Volume do jornal/revista
stg	textos_jornais	issn	varchar(20)	International Standard Serial Number
stg	textos_jornais	autores	text	Lista de autores do texto
stg	trabalhos_eventos	id	serial(4)	Identificador único do trabalho
stg	trabalhos_eventos	id_lattes	varchar(100)	Identificador do pesquisador no Lattes
stg	trabalhos_eventos	ano	varchar(10)	Ano de publicação do trabalho
stg	trabalhos_eventos	titulo	varchar(1000)	Título do trabalho apresentado
stg	trabalhos_eventos	nome_evento	varchar(500)	Nome do evento científico
stg	trabalhos_eventos	titulo_anais	varchar(500)	Título dos anais do evento
stg	trabalhos_eventos	doi	varchar(255)	Digital Object Identifier do trabalho
stg	trabalhos_eventos	idioma	varchar(50)	Idioma de publicação
stg	trabalhos_eventos	natureza	varchar(50)	Natureza do trabalho (resumo/completo)
stg	trabalhos_eventos	meio_divulgacao	varchar(50)	Meio de divulgação (impresso/digital)
stg	trabalhos_eventos	pais_evento	varchar(100)	País onde o evento foi realizado
stg	trabalhos_eventos	ano_realizacao	varchar(10)	Ano de realização do evento

stg	trabalhos_eventos	cidade_evento	varchar(255)	Cidade onde o evento foi realizado
stg	trabalhos_eventos	classificação_evento	varchar(50)	Classificação do evento (nacional/internacional)
stg	trabalhos_eventos	nome_editora	varchar(255)	Nome da editora dos anais
stg	trabalhos_eventos	cidade_editora	varchar(255)	Cidade da editora
stg	trabalhos_eventos	isbn	varchar(50)	International Standard Book Number dos anais
stg	trabalhos_eventos	volume	varchar(20)	Volume dos anais
stg	trabalhos_eventos	pagina_inicial	varchar(10)	Número da página inicial
stg	trabalhos_eventos	pagina_final	varchar(10)	Número da página final
stg	trabalhos_eventos	autores	text	Lista de autores do trabalho
stg	apresentações_trabalho	id	serial(4)	Identificador único da apresentação
stg	apresentações_trabalho	id_lattes	varchar(100)	Identificador do pesquisador no Lattes
stg	apresentações_trabalho	ano	varchar(10)	Ano da apresentação
stg	apresentações_trabalho	titulo	varchar(1000)	Título do trabalho apresentado
stg	apresentações_trabalho	doi	varchar(255)	Digital Object Identifier da apresentação
stg	apresentações_trabalho	idioma	varchar(50)	Idioma da apresentação
stg	apresentações_trabalho	natureza	varchar(100)	Natureza da apresentação (palestra/poster/oral)
stg	apresentações_trabalho	pais	varchar(100)	País onde foi realizada a apresentação
stg	apresentações_trabalho	nome_evento	varchar(500)	Nome do evento
stg	apresentações_trabalho	cidade_apresentacao	varchar(255)	Cidade da apresentação
stg	apresentações_trabalho	local_apresentacao	varchar(255)	Local específico da apresentação
stg	apresentações_trabalho	instituição_promotora	varchar(500)	Instituição organizadora do evento
stg	apresentações_trabalho	autores	text	Lista de autores do trabalho
stg	outras_producoes	id	serial(4)	Identificador único da produção

stg	outras_producoes	id_lattes	varchar(100)	Identificador do pesquisador no Lattes
stg	outras_producoes	ano	varchar(10)	Ano da produção
stg	outras_producoes	titulo	varchar(1000)	Título da produção
stg	outras_producoes	doi	varchar(255)	Digital Object Identifier da produção
stg	outras_producoes	idioma	varchar(50)	Idioma da produção
stg	outras_producoes	natureza	varchar(200)	Natureza ou tipo da produção
stg	outras_producoes	meio_divulgacao	varchar(50)	Meio de divulgação (impresso/digital)
stg	outras_producoes	pais_publicacao	varchar(100)	País de publicação
stg	outras_producoes	cidade_editora	varchar(255)	Cidade da editora
stg	outras_producoes	editora	varchar(500)	Nome da editora
stg	outras_producoes	issn_isbn	varchar(50)	ISSN ou ISBN da publicação
stg	outras_producoes	numero_paginas	varchar(20)	Número de páginas
stg	outras_producoes	autores	text	Lista de autores
dw	dim_area	id_area	serial(4)	Identificador único da área
dw	dim_area	grande_area	varchar(255)	Nome da grande área do conhecimento
dw	dim_area	area	varchar(255)	Nome da área específica
dw	dim_linha_pesquisa	id_linha_pesquisa	serial(4)	Identificador único da linha de pesquisa
dw	dim_linha_pesquisa	linha_pesquisa	varchar(500)	Nome normalizado da linha de pesquisa
dw	dim_pesquisador	id_pesquisador	serial(4)	Identificador único do pesquisador
dw	dim_pesquisador	id_lattes	varchar(100)	Identificador do pesquisador no currículo Lattes
dw	dim_pesquisador	nome	varchar(100)	Nome completo do pesquisador
dw	dim_pesquisador	atuacao_profissional	text	Instituição de atuação profissional do pesquisador
dw	dim_tempo	id_tempo	serial(4)	Identificador único do ano

dw	dim_tempo	ano	(int4)	Ano referente à dimensão temporal
dw	dim_tipo_producao	id_tipo_producao	serial(4)	Identificador único do tipo de produção (chave primária - gerado automaticamente)
dw	dim_tipo_producao	tipo_producao	varchar(100)	Tipo de produção científica (Artigos/Livros/Trabalhos em Eventos/etc)
dw	dim_localizacao_trabalhos	id_localizacao_trabalhos	serial(4)	Identificador único da localização
dw	dim_localizacao_trabalhos	id_lattes	varchar(100)	Identificador do pesquisador no Lattes
dw	dim_localizacao_trabalhos	instituicao	varchar(500)	Nome da instituição relacionada à produção
dw	dim_localizacao_trabalhos	pais	varchar(100)	País relacionado à produção
dw	fato_pesquisador_producoes	id_fato_pesquisador_producoes	serial(4)	Identificador único do registro de fato
dw	fato_pesquisador_producoes	id_pesquisador	(int4)	Identificador do pesquisador
dw	fato_pesquisador_producoes	id_area	(int4)	Identificador do ano
dw	fato_pesquisador_producoes	id_tipo_producao	(int4)	Identificador do tipo de produção
dw	fato_pesquisador_producoes	qtd_producoes	(int4)	Quantidade de produções do pesquisador no ano e tipo especificados
dw	fato_pesquisador_area_atuacao	id_fato_pesquisador_area_atuacao	serial(4)	Identificador único do registro de fato
dw	fato_pesquisador_area_atuacao	id_pesquisador	(int4)	Identificador do pesquisador
dw	fato_pesquisador_area_atuacao	id_area	(int4)	Identificador da área
dw	fato_pesquisador_area_atuacao	presenca	(int4)	Indicador de presença (1 = pesquisador atua nesta área)
dw	fato_pesquisador_linha_pesquisa	id_fato_pesquisador_linha	serial(4)	Identificador único do registro de fato
dw	fato_pesquisador_linha_pesquisa	id_pesquisador	(int4)	Identificador do pesquisador
dw	fato_pesquisador_linha_pesquisa	id_linha_pesquisa	(int4)	Identificador da linha de pesquisa
dw	fato_pesquisador_linha_pesquisa	presenca	(int4)	Indicador de presença (1 = pesquisador atua nesta linha)

dw	fato_pesquisador_producao_localizacao	id_fato_pesq_prod_loc	serial(4)	Identificador único do registro de fato
dw	fato_pesquisador_producao_localizacao	id_pesquisador	(int4)	Identificador do pesquisador
dw	fato_pesquisador_producao_localizacao	id_tempo	(int4)	Identificador do ano
dw	fato_pesquisador_producao_localizacao	id_tipo_producao	(int4)	Identificador do tipo de produção
dw	fato_pesquisador_producao_localizacao	id_localizacao_trabalhos	(int4)	Identificador da localização
dw	fato_pesquisador_producao_localizacao	qtd_producoes	(int4)	Quantidade de produções do pesquisador por localização