INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
DE MINAS GERAIS (IFMG)
CAMPUS BAMBUÍ
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

Vinícius Tadeu Andrade Costa

AUTOMAÇÃO DE COLETA E ESTRUTURAÇÃO DE DADOS DA CVM PARA ANÁLISE FUNDAMENTALISTA

Bambuí – MG
2025

VINÍCIUS TADEU ANDRADE COSTA

AUTOMAÇÃO DE COLETA E ESTRUTURAÇÃO DE DADOS DA CVM PARA ANÁLISE FUNDAMENTALISTA

Trabalho de conclusão de curso apresentado ao Curso de Bacharelado em Engenharia de Computação do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG) – *Campus* Bambuí para obtenção do grau de Bacharel em Engenharia de Computação.

Orientador: Prof. Dr. Marcos Roberto Ribeiro

Bambuí – MG
2025

Catalogação na Fonte Biblioteca IFMG - Campus Bambuí

C837a Costa, Vinícius Tadeu Andrade.
Automação de coleta e estruturação de dados da CVM para análise fundamentalista. / Vinícius Tadeu Andrade Costa. – 2025.
84 f.; il.: color.

Orientador: Prof. Dr. Marcos Roberto Ribeiro.
Trabalho de Conclusão de Curso (graduação) - Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – Campus Bambuí, MG, Curso Bacharelado em Engenharia de Produção, 2025.

1. CVM. 2. Análise fundamentalista. 3. Dados financeiros. I. Ribeiro, Marcos Roberto. II. Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – Campus Bambuí, MG. III. Título.

CDD 005.12

Elaborada por Douglas Bernardes de Castro- CRB-6/2802

Vinícius Tadeu Andrade Costa

AUTOMAÇÃO DE COLETA E ESTRUTURAÇÃO DE DADOS DA CVM PARA ANÁLISE FUNDAMENTALISTA

Trabalho de conclusão de curso apresentado ao Curso de Bacharelado em Engenharia de Computação do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG) – *Campus* Bambuí para obtenção do grau de Bacharel em Engenharia de Computação.

Aprovado em 01 de Agosto de 2025 pela banca examinadora:

Prof. Dr. Marcos Roberto Ribeiro – IFMG – *Campus* Bambuí – (Orientador)
Prof. Me. Álvaro Antônio Fonseca de Souza – IFMG - *Campus* Bambuí
Prof. Me. Cláudio Ribeiro de Sousa – IFMG - *Campus* Bambuí

Documento assinado eletronicamente por **Marcos Roberto Ribeiro**, **Professor**, em 01/08/2025, às 14:36, conforme Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Álvaro Antonio Fonseca de Souza**, **Professor EBTT**, em 01/08/2025, às 14:47, conforme Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Claudio Ribeiro de Sousa**, **Professor EBTT**, em 01/08/2025, às 14:47, conforme Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade do documento pode ser conferida no site https://sei.ifmg.edu.br/consultadocs informando o código verificador **2382834** e o código CRC **6A9A9BB0**.

Dedico este trabalho à minha esposa, minha principal fonte de amor, apoio e inspiração. Seu incentivo constante e sua fé inabalável foram fundamentais para que eu mantivesse a confiança em meus objetivos, mesmo diante dos desafios.

AGRADECIMENTOS

Agradeço profundamente à minha família — minha esposa, meus pais, meu irmão, minha sogra e meu sogro — por todo o apoio, paciência e palavras de encorajamento ao longo desta jornada. A presença e o carinho de vocês foram essenciais para que eu não desistisse e continuasse acreditando na concretização deste objetivo.

Agradeço, ainda, ao meu orientador, pela paciência, orientação e valiosas contribuições para o desenvolvimento deste trabalho.

“Faça o teu melhor, na condição que você tem, enquanto você não tem condições melhores para fazer melhor ainda.”

Mario Sergio Cortella

RESUMO

Este trabalho apresenta o desenvolvimento de um *software* automatizado para coleta, estruturação e integração de dados financeiros públicos disponibilizados pela Comissão de Valores Mobiliários (CVM), com foco na aplicação da análise fundamentalista. A solução proposta visa superar as limitações de acesso, padronização e organização dos dados financeiros divulgados por companhias abertas, que tradicionalmente são apresentados em formatos técnicos e fragmentados. O *software* foi desenvolvido em Python e utiliza uma arquitetura baseada em etapas de extração, transformação e carga, apoiada por um modelo relacional implementado em SQLite, um sistema de gerenciamento de banco de dados relacional leve, embutido e de fácil integração com aplicações Python. A ferramenta permite integrar diversas categorias de dados da CVM, como demonstrações financeiras padronizadas, informações trimestrais, formulários cadastrais e comunicados relevantes. Após o processamento, esses dados estruturados foram utilizados para cálculo de indicadores fundamentalistas, demonstrando a aplicabilidade prática da solução proposta. O projeto contribui com a democratização da informação financeira, promovendo maior transparência, acessibilidade e reprodutibilidade de análises no mercado financeiro. Além disso, oferece suporte à pesquisa acadêmica, desenvolvimento de ferramentas quantitativas e aplicações educacionais na área de finanças.

Palavras-chave: CVM. Análise Fundamentalista. Dados Financeiros. Banco de dados.

ABSTRACT

This work presents the development of an automated *software* for the collection, structuring, and integration of public financial data provided by the Brazilian Securities and Exchange Commission (CVM), focusing on the application of fundamental analysis. The proposed solution aims to overcome the limitations of access, standardization, and organization of financial data disclosed by publicly traded companies, which are traditionally presented in technical and fragmented formats. The *software* was developed in Python and uses an architecture based on extraction, transformation, and loading stages, supported by a relational model implemented in SQLite, a lightweight, embedded, and easily integrable relational database management system for Python applications. The tool allows the integration of various CVM data categories, such as standardized financial statements, quarterly information, registration forms, and relevant announcements. After processing, these structured data were used to calculate fundamental indicators, demonstrating the practical applicability of the proposed solution. The project contributes to the democratization of financial information, promoting greater transparency, accessibility, and reproducibility of analyses in the financial market. In addition, it supports academic research, the development of quantitative tools, and educational applications in the field of finance.

Keywords: CVM. Fundamental Analysis. Financial Data. Database.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS

B3 – Brasil, Bolsa, Balcão
CSV – *Comma-Separated Values* (Valores Separados por Vírgula)
CVM – Comissão de Valores Mobiliários
DCF – *Discounted Cash Flow* (Fluxo de Caixa Descontado)
DER – Diagrama Entidade-Relacionamento
DFP – Demonstrações Financeiras Padronizadas
DY – *Dividend Yield*
ETL – *Extract, Transform, Load* (Extração, Transformação e Carga)
FCA – Formulário Cadastral
FRE – Formulário de Referência
IAN – Informações Anuais
IPO – *Initial Public Offering* (Oferta Pública Inicial)
ITR – Informações Trimestrais
LPA – Lucro por Ação
PDF – *Portable Document Format*
P/L – *Price-to-Earnings Ratio* (Preço por Lucro)
SGBD – Sistema de Gerenciamento de Banco de Dados
SQL – *Structured Query Language*
XML – *Extensible Markup Language*
XBRL – *eXtensible Business Reporting Language*

SUMÁRIO

1 INTRODUÇÃO

O mercado financeiro brasileiro tem passado por transformações significativas nas últimas décadas, impulsionadas pelo avanço tecnológico, pelo aumento do número de investidores e pela necessidade crescente de transparência e acessibilidade das informações financeiras. Segundo Santos (2023), essas mudanças refletem uma busca por maior eficiência e modernização no acesso aos dados do setor. Em especial, o uso da análise fundamentalista como ferramenta para embasar decisões de investimento tem ganhado destaque, exigindo dados estruturados, confiáveis e de fácil acesso (DANTAS, 2020).

Esse movimento reflete a busca constante por maior eficiência e acessibilidade nas informações financeiras, necessárias para investidores e analistas que utilizam a análise fundamentalista como base para a tomada de decisões. O crescimento desse interesse pode ser observado no relatório anual de 2023 da Brasil, Bolsa e Balcão (B3), que apontou um aumento de aproximadamente 15% no número de investidores em comparação com 2022 (B3, 2023c). Esse aumento demonstra a importância de facilitar o acesso aos dados empresariais para ampliar a capacidade de análise e decisão no mercado.

No Brasil, dois órgãos desempenham papéis centrais na disponibilização de dados do mercado de capitais: a B3 e a Comissão de Valores Mobiliários (CVM). A B3, fundada em 1890 e sediada em São Paulo, é a única bolsa de valores do país e fornece dados sobre o histórico de negociações de ativos (B3, 2023b). Já a CVM, criada em 1976, é a entidade responsável por regulamentar e supervisionar o mercado de capitais brasileiro, disponibilizando publicamente uma ampla gama de informações contábeis e financeiras das empresas listadas na bolsa (CVM, 2009).

Embora esses órgãos disponibilizem dados de forma pública, as informações encontram-se frequentemente dispersas, em formatos técnicos e de difícil manipulação, especialmente para usuários sem formação avançada em ciência de dados ou tecnologia da informação. A CVM, por exemplo, oferece demonstrações financeiras, formulários periódicos, informes de governança e dados cadastrais das companhias abertas. No entanto, a ausência de integração entre essas fontes e a falta de padronização dificultam análises sistemáticas e a construção de séries históricas coerentes.

Tais limitações representam barreiras importantes para analistas, investidores e pesquisadores que desejam realizar estudos mais aprofundados sobre o desempenho das companhias abertas brasileiras. A dificuldade de manipular os dados restringe o acesso à informação e compromete a capacidade analítica de boa parte do mercado, mostrando a necessidade de ferramentas que automatizem a coleta, organização e disponibilização desses dados em um formato mais acessível e estruturado.

A necessidade de soluções mais acessíveis já foi abordada por estudos acadêmicos. Por exemplo, Guilherme e Marotti (2021) propõem um modelo de dados flexível para análise fundamentalista moderna, estruturando balanço patrimonial, demonstração de resultados e fluxo de caixa em um banco de dados não relacional. Essa proposta reforça o valor de abordagens que promovam a organização dos dados financeiros de maneira simplificada e consistente.

Diante desse contexto, este trabalho propôs o desenvolvimento de um sistema automatizado para a integração, coleta, estruturação e análise dos dados financeiros públicos fornecidos pela CVM, facilitando seu acesso e processamento. A solução visa oferecer uma base de dados unificada, capaz de extrair e organizar essas informações de forma eficiente, contribuindo para a análise fundamentalista e auxiliando na tomada de decisões.

1.1 Objetivos

O objetivo principal deste trabalho foi desenvolver um sistema automatizado para integrar e estruturar os dados públicos da CVM, possibilitando sua utilização em análises fundamentalistas. Para isso, foram definidos os seguintes objetivos específicos:

- analisar a estrutura e os padrões dos conjuntos de dados disponibilizados pela CVM;
- projetar e implementar um modelo relacional conforme as boas práticas de modelagem de dados, visando ao desempenho e à consistência;
- desenvolver um processo automatizado de coleta, tratamento e carga dos dados em banco relacional;
- demonstrar a aplicação prática da base estruturada na construção de indicadores fundamentais.

1.2 Justificativa

O mercado de capitais brasileiro vem registrando crescimento expressivo nos últimos anos, tanto em volume de negociações quanto na participação de investidores. De acordo com dados da B3, em 2023, o volume total negociado alcançou R$ 7,2 trilhões, representando um aumento de 12,6% em relação a 2022. No mesmo período, o capital efetivamente movimentado foi de R$ 2,4 trilhões, indicando um crescimento de 15,5% (B3, 2023b).

Esse cenário reflete não apenas o amadurecimento do mercado, mas também o aumento do interesse de pessoas físicas na bolsa de valores, tornando cada vez mais necessária a disponibilidade de ferramentas que simplifiquem o acesso à

informação e apoiem decisões fundamentadas.

Embora os dados da B3 não sejam diretamente utilizados neste trabalho, seu contexto é relevante por representar o principal ambiente de negociação de ativos financeiros no País. As informações que subsidiam essas negociações, como demonstrações contábeis, eventos relevantes e cadastros de companhias, são fornecidas pela CVM, órgão responsável por regular e fiscalizar o mercado de capitais.

Entretanto, os conjuntos de dados fornecidos pela CVM apresentam limitações significativas: são fragmentados, distribuídos em diferentes formatos e fontes e demandam elevado grau de conhecimento técnico para extração e análise. Essa complexidade dificulta o uso sistemático e automatizado dos dados por parte de analistas, investidores e pesquisadores.

Até o momento, não foram identificados trabalhos que promovam uma integração estruturada, sistemática e publicamente acessível dos dados disponibilizados pela CVM. Assim, este trabalho se justifica pela proposta de desenvolver uma solução automatizada que integre essas informações em uma base de dados relacional organizada, promovendo maior acessibilidade, eficiência analítica e reprodutibilidade científica.

1.3 Estrutura do documento

Este trabalho está organizado da seguinte maneira: o Capítulo 2 discute os fundamentos teóricos que embasam a pesquisa, com ênfase em conceitos relacionados ao mercado financeiro, análise fundamentalista e modelagem de dados. O Capítulo 3 apresenta a metodologia adotada, detalhando a classificação do trabalho, a abordagem utilizada e a organização das etapas do projeto. Em seguida, o Capítulo 4 descreve o processo de desenvolvimento do sistema proposto, incluindo a análise dos dados da CVM, a modelagem relacional e a implementação de um processo automatizado de coleta e estruturação dos dados. O Capítulo 5 expõe os resultados obtidos, destaca as contribuições e limitações da solução desenvolvida e propõe direções para trabalhos futuros. Por fim, são apresentadas as considerações finais.

2 FUNDAMENTOS TEÓRICOS

Este capítulo apresenta os fundamentos teóricos que sustentam o desenvolvimento deste trabalho, bem como os principais estudos e iniciativas relacionadas ao tema. A Seção 2.1 aborda o mercado financeiro brasileiro, com ênfase na atuação da CVM e sua função reguladora. A Seção 2.2 discute as diferentes abordagens de análise para investimentos, com foco na análise fundamentalista. A Seção 2.3 descreve a modelagem dos dados e os processos de integração utilizados no projeto. Por fim, a Seção 2.4 apresenta trabalhos acadêmicos e iniciativas similares à proposta deste estudo, destacando seus principais diferenciais.

2.1 Mercado financeiro

O sistema financeiro é formado por um conjunto de instituições, instrumentos e operações voltados à administração dos recursos econômicos. Ele se divide em diferentes áreas, como o mercado de capitais, o monetário, o cambial e o de derivativos, tendo como função principal intermediar a transferência de recursos entre agentes superavitários e deficitários, além de oferecer alternativas de investimento e financiamento para empresas, governos e indivíduos (TEIXEIRA, 2019; DAMODARAN, 2012).

No contexto brasileiro, essa estrutura é segmentada em cinco principais componentes: capitais, monetário, crédito, câmbio e derivativos. O segmento de capitais é responsável pela negociação de ações, debêntures e outros ativos mobiliários, proporcionando às empresas a captação de recursos por meio da emissão de títulos negociáveis (SUNO, 2021).

O setor monetário, por sua vez, abrange operações de curto prazo realizadas entre instituições financeiras e o Banco Central, com o propósito de regular a liquidez do sistema e controlar as taxas de juros. Essa dinâmica contribui para a manutenção do equilíbrio macroeconômico (SOUZA FIGUEIREDO *et al.*, 2021).

O componente de crédito engloba atividades destinadas ao financiamento de pessoas físicas e jurídicas, por meio de empréstimos e financiamentos concedidos por instituições financeiras. Trata-se de um mecanismo essencial para fomentar o consumo, viabilizar investimentos e impulsionar o crescimento econômico.

As operações relacionadas à troca de moedas estrangeiras estão inseridas no segmento cambial, cuja finalidade é viabilizar o comércio internacional, atrair investimentos externos e permitir a adequada gestão das reservas cambiais nacionais (GÓIS; SOARES, 2019).

O segmento de derivativos contempla instrumentos como contratos futuros, opções e *swaps*, utilizados como mecanismos de proteção contra riscos financeiros.

Essa estrutura permite que agentes econômicos se antecipem a oscilações adversas em variáveis como preços de ativos e taxas de juros (FIGUEIREDO, 2023).

Dentro do mercado financeiro, as ações representam uma das principais modalidades de investimento. Uma ação é uma fração do capital social de uma empresa, conferindo ao seu detentor a propriedade de uma parcela da empresa emissora. Ao adquirir ações, o investidor se torna um sócio da empresa e pode ter direito à participação nos lucros distribuídos e, em alguns casos, a voto nas assembleias de acionistas (SUNO, 2021; GOMES, 2007).

A negociação de ações ocorre principalmente em bolsas de valores, como a B3 no Brasil, onde investidores podem comprar e vender ativos. O preço das ações é determinado pela oferta e demanda, ou seja, pela interação entre compradores e vendedores no mercado (SANTANDER, 2024). Além disso, diversos fatores influenciam a valorização ou desvalorização dos papéis, como os resultados financeiros das empresas, a conjuntura econômica e eventos geopolíticos (DAMODARAN, 2012; ATTIE, 2013).

Duas instituições fundamentais nesse ecossistema são a B3 e a CVM. A B3 atua como principal plataforma de negociação de ativos financeiros, enquanto a CVM desempenha um papel regulador, garantindo transparência e equidade no mercado de capitais (CVM, 2023a).

Criada pela Lei n.º 6.385/1976 (BRASIL, 1976a), a CVM tem como principal missão proteger os investidores e garantir o funcionamento eficiente do mercado de capitais (CVM, 2009). Entre suas principais atribuições, estão (CVM, 2009; ROBERTO, 2023):

- regulamentação da abertura de capital de empresas e suas obrigações com o mercado;
- supervisão de corretoras, bancos de investimento e outras instituições financeiras;
- fiscalização da negociação de ativos financeiros e combate a fraudes e manipulações de mercado;
- garantia da transparência e divulgação de informações pelas companhias de capital aberto.

A CVM exige que todas as empresas listadas na bolsa de valores publiquem periodicamente suas demonstrações financeiras e outros documentos essenciais, como balanços patrimoniais e informações sobre acionistas relevantes. Essa transparência fortalece a confiança dos investidores e contribui para a estabilidade e o crescimento do mercado financeiro brasileiro (FGV, 2024).

Para garantir a disponibilidade dessas informações, a CVM mantém uma base de dados acessível ao público por meio de seu portal eletrônico[1] e sistemas es-

[1] https://dados.cvm.gov.br.

pecializados, como o Sistema Empresas.NET e o CVMWeb. As empresas registradas são obrigadas a enviar seus documentos periodicamente, os quais são disponibilizados em diversos formatos.

A base de dados da CVM inclui:

- demonstrações financeiras padronizadas (DFPs);
- informes trimestrais (ITRs);
- relatórios de governança corporativa;
- documentos de ofertas públicas de ações (IPO);
- dados sobre fundos de investimento e gestores de ativos.

2.2 Análises de investimento

A avaliação de ativos financeiros exige metodologias analíticas capazes de mensurar riscos, retornos esperados e fundamentos econômicos. Nesse contexto, distintas abordagens são utilizadas para subsidiar decisões de alocação de capital, entre as quais se destacam a análise técnica, a análise de sentimento, a análise quantitativa, a análise macroeconômica e a análise fundamentalista (LIAW, 2011).

A análise técnica concentra-se no comportamento histórico dos preços dos ativos e volumes de negociação, com o objetivo de identificar padrões recorrentes que possam indicar movimentos futuros. Essa abordagem busca antecipar pontos de reversão, suportes e resistências, sendo amplamente empregada por operadores de curto prazo (MAURICE; AGYARKO; PAUL, 2019).

A análise de sentimento, por sua vez, considera as percepções e expectativas dos agentes de mercado, captadas a partir de fontes como notícias, mídias sociais, fóruns e relatórios de analistas. Técnicas de processamento de linguagem natural e modelos de aprendizado de máquina são aplicados para extrair métricas subjetivas do comportamento coletivo, permitindo a construção de indicadores qualitativos com impacto sobre os preços dos ativos (KEARNEY; LIU, 2014).

No campo quantitativo, predominam modelos estatísticos e computacionais voltados à identificação de padrões e relações complexas entre variáveis financeiras. Regressões, séries temporais, redes neurais e algoritmos genéticos figuram entre os recursos empregados para construir estratégias de investimento baseadas em evidência empírica e análise de grandes volumes de dados (SAHU; MOKHADE; BOKDE, 2023).

A abordagem macroeconômica, por sua natureza abrangente, investiga variáveis estruturais que influenciam o ambiente de mercado como um todo. Indicadores como inflação, taxa de juros, produto interno bruto (PIB) e política fiscal e monetária são analisados com o intuito de compreender os ciclos econômicos e antecipar movimentos sistêmicos que impactam diversos setores simultaneamente (CLAESSENS;

KOSE, 2017).

Entre as metodologias mencionadas, a análise fundamentalista assume papel central neste estudo, por sua capacidade de mensurar o valor econômico intrínseco dos ativos com base em fundamentos financeiros e operacionais. Essa abordagem sustenta-se na hipótese de que o valor de mercado pode divergir temporariamente do valor real de um ativo, sendo possível identificar oportunidades de investimento por meio dessa assimetria informacional (JOBIM, 2025).

O valor intrínseco corresponde a uma estimativa do verdadeiro valor de uma empresa ou ativo, calculado com base em elementos como receitas, lucros, geração de caixa, estrutura de capital, posição competitiva e governança corporativa. Tal valor reflete o potencial da empresa em gerar resultados sustentáveis ao longo do tempo, independentemente das flutuações de mercado (DEREK *et al.*, 2025).

Para conduzir essa estimativa, a análise fundamentalista recorre a diversos indicadores que mensuram aspectos distintos da situação financeira empresarial. Entre os principais grupos, estão: indicadores de rentabilidade, que avaliam a capacidade de geração de lucro; indicadores de liquidez, voltados à solvência de curto prazo; indicadores de endividamento, que mensuram o grau de alavancagem; e métricas de avaliação de valor, como o fluxo de caixa descontado e múltiplos de mercado (BABALOLA; ABIOLA, 2013).

2.2.1 Indicadores de rentabilidade

Os indicadores de rentabilidade medem a eficiência da empresa em gerar lucros em relação a diferentes bases financeiras. Esses cálculos são amplamente utilizados em estudos financeiros e derivam da contabilidade gerencial e da análise de demonstrações financeiras (KOTHARI, 2001).

O lucro por ação (LPA) indica o montante de lucro líquido atribuído a cada ação ordinária em circulação da empresa. É um indicador usado para avaliar o desempenho das ações no mercado. Sua fórmula é dada pela Equação (2.1a). Considerando um lucro líquido de R$ 10 milhões e 2 milhões de ações em circulação, tem-se o LPA de R$ 5,00, conforme mostrado na Equação (2.1b) (MAURICE; AGYARKO; PAUL, 2019).

$$\text{LPA} = \frac{\text{Lucro líquido}}{\text{Número de ações}} \qquad \text{(2.1a)} \qquad \text{LPA} = \frac{10.000.000{,}00}{2.000.000{,}00} = 5{,}00 \qquad \text{(2.1b)}$$

O índice preço por lucro (P/L) mede quanto os investidores estão dispostos a pagar pelo lucro da empresa. Sua fórmula está na Equação (2.2a). Como exemplo, se uma ação custa R$ 50,00 e o LPA é R$ 5,00, o P/L será 10,00, indicando

que os investidores estão dispostos a pagar 10 vezes o lucro anual por ação (SAHU; MOKHADE; BOKDE, 2023).

$$\mathrm{P/L} = \frac{\text{Preço da ação}}{\text{LPA}} \qquad \text{(2.2a)} \qquad\qquad \mathrm{P/L} = \frac{50{,}00}{5{,}00} = 10{,}00 \qquad \text{(2.2b)}$$

O *dividend yield* (DY) mede a rentabilidade dos dividendos pagos em relação ao preço da ação, conforme Equação (2.3a). Se a empresa paga R$ 2,00 de dividendos por ação, e o preço da ação é R$ 40,00, o cálculo da Equação (2.3b) resulta em um DY de 5,00% (KEARNEY; LIU, 2014).

$$\mathrm{DY} = \frac{\text{Dividendo por ação}}{\text{Preço da ação}} \times 100 \quad \text{(2.3a)} \qquad\qquad \mathrm{DY} = \frac{2{,}00}{40{,}00} \times 100 = 5{,}00\% \qquad \text{(2.3b)}$$

2.2.2 Indicadores de liquidez

Os indicadores de liquidez avaliam a capacidade da empresa de honrar suas obrigações de curto prazo (CLAESSENS; KOSE, 2017). Esses indicadores comparam recursos disponíveis com dívidas exigíveis no curto prazo, destacando a importância dos conceitos de ativo circulante e passivo circulante.

O ativo circulante representa os bens e direitos que a empresa espera realizar, vender ou consumir no curso normal de suas operações dentro de um período de até doze meses. Isso inclui disponibilidades, contas a receber, estoques e aplicações financeiras de liquidez imediata.

Já o passivo circulante compreende as obrigações que a entidade deve liquidar nesse mesmo período, como fornecedores, salários a pagar, tributos e empréstimos de curto prazo. A correta gestão desses elementos é essencial para garantir o equilíbrio financeiro da organização.

A liquidez corrente (LC), por sua vez, é um dos principais indicadores utilizados para mensurar essa capacidade de pagamento no curto prazo. Ela é expressa na Equação (2.4a), sendo a razão entre o ativo circulante e o passivo circulante. Se o ativo circulante é de R$ 500.000, e o passivo circulante, de R$ 250.000, a liquidez corrente é igual a 2,00, conforme a Equação (2.4b).

$$\mathrm{LC} = \frac{\text{Ativo circulante}}{\text{Passivo circulante}} \qquad \text{(2.4a)} \qquad\qquad \mathrm{LC} = \frac{500.000}{250.000} = 2{,}00 \qquad \text{(2.4b)}$$

2.2.3 Indicadores de endividamento

Os indicadores de endividamento mostram a dependência da empresa em relação ao capital de terceiros (KOTHARI, 2001). Em geral, quanto maior a proporção

de dívidas em relação ao capital próprio, maior é o risco financeiro do negócio.

A análise começa pela compreensão do passivo não circulante, que representa as obrigações exigíveis com prazo de vencimento superior a 12 meses. Dentre os principais exemplos, estão os financiamentos de longo prazo, debêntures, empréstimos bancários com vencimento além do exercício seguinte e provisões de longo prazo.

Esses elementos estão associados a estratégias de financiamento estruturado, investimentos em ativos de longo prazo ou reorganização da estrutura de capital da empresa. Conforme a Lei n.º 6.404/76, art. 179, o passivo não circulante abrange as obrigações com vencimento após o término do exercício social seguinte (BRASIL, 1976b).

Um dos principais indicadores utilizados é a dívida bruta, que representa a soma do passivo circulante (obrigações de curto prazo) e do passivo não circulante (obrigações de longo prazo). A Equação (2.5a) expressa a fórmula geral da dívida bruta. A seguir, a Equação (2.5b) apresenta um exemplo numérico no qual a dívida bruta totaliza R$ 3.000.000, resultante da soma de R$ 1.000.000 de passivo circulante e R$ 2.000.000 de passivo não circulante.

$$\text{Dívida bruta} = \text{Passivo circulante} + \text{Passivo não circulante} \tag{2.5a}$$

$$\text{Dívida bruta} = 1.000.000 + 2.000.000 = 3.000.000 \tag{2.5b}$$

Contudo, nem toda a dívida bruta representa, de fato, uma exigência líquida de recursos. Por isso, utiliza-se, também, o indicador de dívida líquida, que representa a real necessidade de capital de terceiros, já que desconta os valores disponíveis em caixa ou aplicados com liquidez imediata. A dívida bruta subtraída dos recursos que podem ser prontamente utilizados pela empresa. A Equação (2.6a) define essa relação. No exemplo da Equação (2.6b), ao subtrair R$ 500.000 de caixa e equivalentes de caixa de uma dívida bruta de R$ 3.000.000, obtém-se uma dívida líquida de R$ 2.500.000.

$$\text{Dívida líquida} = \text{Dívida bruta} - \text{Caixa e equivalentes de caixa} \tag{2.6a}$$

$$\text{Dívida líquida} = 3.000.000 - 500.000 = 2.500.000 \tag{2.6b}$$

2.2.4 Valor intrínseco

Conforme discutido anteriormente, o valor intrínseco representa uma estimativa do verdadeiro valor econômico de um ativo, calculado com base em seus

fundamentos financeiros e capacidade futura de geração de resultados. Para operacionalizar essa estimativa, a análise fundamentalista, frequentemente, recorre ao modelo de desconto de fluxo de caixa (DCF), que quantifica o valor presente dos fluxos de caixa futuros esperados. Esse método parte do princípio de que o valor de um ativo corresponde à soma dos fluxos de caixa futuros, trazidos ao valor presente por meio de uma taxa de desconto apropriada.

O modelo DCF é representado na Equação (2.7a), na qual r indica a taxa de desconto, e t , o período de tempo. Com um fluxo de caixa de R$ 1.000.000 projetado para cinco anos no futuro e taxa de desconto de 10%, o valor intrínseco resulta em R$ 620.921,32, conforme demonstrado na Equação (2.7b). Esse cálculo reflete o valor presente de um fluxo de caixa futuro a ser recebido em cinco anos (FIGUEIREDO, 2023).

$$\text{Valor intrínseco} = \sum \frac{\text{Fluxo de caixa esperado}}{(1+r)^t} \tag{2.7a}$$

$$\text{Valor intrínseco} = \frac{1.000.000}{(1+0{,}10)^5} = \frac{1.000.000}{1{,}610.51} = 620.921{,}32 \tag{2.7b}$$

2.3 Modelagem e integração de dados

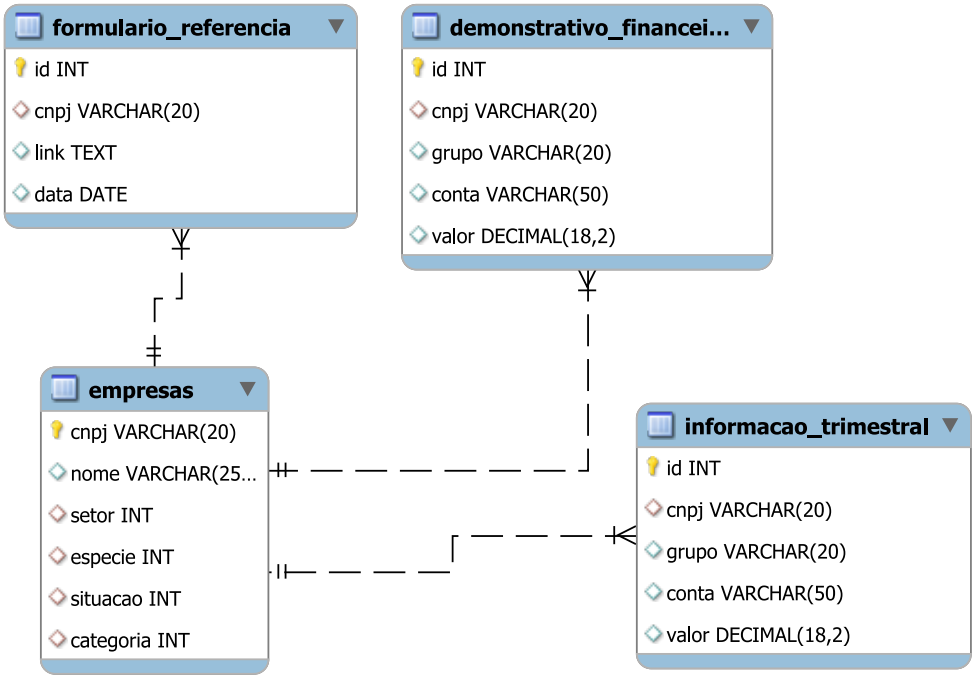
A modelagem de dados constitui uma etapa fundamental no desenvolvimento de sistemas de informação, especialmente em contextos caracterizados por volumes elevados de dados e por complexidade estrutural significativa, como é frequente no setor financeiro. Esse procedimento refere-se à representação estruturada dos elementos essenciais de determinado domínio, abrangendo entidades, atributos e relacionamentos. Tal abordagem tem como propósito viabilizar a organização, a padronização e a manipulação eficiente das informações (ELMASRI; NAVATHE, 2005).

Dentro do processo de modelagem, o nível lógico estabelece a transição entre o modelo conceitual e a implementação física dos dados. O objetivo dessa etapa é converter os requisitos identificados no nível conceitual, geralmente representados por meio de diagramas entidade-relacionamento (DER), em estruturas técnicas formalizadas e compatíveis com sistemas gerenciadores de banco de dados (SGBDs), como MySQL e PostgreSQL (CONNOLLY; BEGG, 2015). Entre os elementos que compõem essa fase, estão tabelas, atributos, tipos de dados, chaves primárias e estrangeiras, além de restrições de integridade e regras para normalização.

A modelagem lógica inclui definições técnicas que determinam a maneira pela qual os dados serão organizados, armazenados e manipulados no ambiente computacional, mantendo correspondência direta com os modelos conceituais, cuja função é representar semanticamente o domínio estudado.

A Figura 1 apresenta um exemplo ilustrativo de esquema lógico de dados, indicando sua estrutura relacional composta por tabelas, relacionamentos e atributos. Esquemas dessa natureza são amplamente empregados para assegurar a organização consistente das informações, permitindo a execução eficiente de consultas complexas e preservando a integridade dos dados ao longo do tempo.

Figura 1 – Exemplo de esquema lógico para banco de dados relacional.



Fonte: Fonte: Elaborado pelo autor (2025).

As tabelas, no modelo lógico, são definidas com o propósito de armazenar dados operacionais e analíticos necessários ao domínio da aplicação. Tipicamente, tais tabelas refletem as entidades identificadas no processo de modelagem conceitual e armazenam atributos como datas, valores monetários, identificadores exclusivos e classificações específicas. A integração dessas tabelas ocorre por meio do uso de chaves primárias e estrangeiras, estabelecendo vínculos coerentes entre os diversos registros presentes no banco de dados.

Tabelas auxiliares podem também compor o modelo lógico, sendo destinadas à organização de domínios controlados e ao suporte à normalização dos dados. Sua utilização contribui para a diminuição de redundâncias e para o aprimoramento da clareza e da integridade das informações armazenadas.

A definição precisa da tipagem dos dados é igualmente fundamental nesse nível de modelagem. Cada atributo recebe um tipo de dado compatível com suas características específicas, garantindo exatidão analítica, eficiência computacional e coerência operacional no contexto dos SGBDs adotados. Entre os tipos mais utilizados, estão *INT*, empregado para valores inteiros, *VARCHAR*, aplicado em textos, e *DECIMAL*, utilizado no armazenamento de valores monetários.

A utilização adequada de uma modelagem lógica estruturada impacta positivamente na qualidade, escalabilidade e desempenho dos sistemas de informação, contribuindo para a robustez, a confiabilidade e a facilidade de manutenção dos sistemas que manipulam dados críticos e complexos.

2.3.1 Processo de integração de dados

A integração de dados é um processo utilizado para combinar informações provenientes de múltiplas fontes heterogêneas, a fim de criar uma visão unificada e consistente que possa ser utilizada por sistemas analíticos, operacionais ou decisórios. Esse processo é aplicado em contextos em que há a necessidade de consolidar dados distribuídos, estruturados ou não, com diferentes formatos, semânticas e origens.

A arquitetura tradicional de integração de dados é baseada no paradigma ETL, sigla para *Extract*, *Transform*, *Load*. Essa abordagem compreende três etapas principais: extração, transformação e carga dos dados em um repositório unificado, geralmente, um banco de dados relacional ou um *data warehouse* (HALEVY; RAJARAMAN; ORDILLE, 2006).

Na etapa de extração, os dados são coletados diretamente das fontes originais. Essas fontes podem incluir bancos de dados transacionais, arquivos estruturados, sistemas legados, APIs, entre outros. A extração deve assegurar a completude, precisão e fidelidade dos dados obtidos, sem comprometer sua integridade.

A etapa de transformação tem como objetivo adaptar os dados extraídos às exigências do sistema de destino. Essa adaptação pode envolver diversas operações, como padronização de formatos, conversão de tipos de dados, limpeza de inconsistências, eliminação de duplicidades, mapeamento de códigos para descrições legíveis e aplicação de regras de negócio. A transformação também pode incluir a reorganização estrutural dos dados por meio da normalização, que busca diminuir redundâncias e dependências, organizando os dados em tabelas de acordo com as formas normais, como a Primeira Forma Normal (1FN), a Segunda Forma Normal (2FN) e a Terceira Forma Normal (3FN) (CODD, 1982).

A normalização é uma técnica aplicada em bancos de dados relacionais que visa decompor tabelas em estruturas menores e mais coesas, com o objetivo de eliminar anomalias de inserção, atualização e exclusão. Esse processo é orientado por regras formais que estabelecem critérios de dependência funcional entre atributos, garantindo integridade e minimizando redundância (CODD, 1982).

Por fim, a etapa de carga consiste na inserção dos dados transformados no ambiente de destino. Essa etapa deve preservar a integridade referencial e garantir que os dados estejam disponíveis de maneira eficiente para consulta e análise. O re-

positório de destino pode adotar diferentes estruturas, como bancos relacionais, data lakes ou sistemas analíticos otimizados.

A aplicação adequada do processo de integração de dados permite a criação de repositórios consistentes, confiáveis e preparados para suportar análises avançadas, extração de conhecimento e processos automatizados de apoio à decisão (RAHM; DO, 2000).

2.4 Referencial Teórico

Nesta seção, são analisados os principais estudos acadêmicos e projetos práticos que fundamentam a abordagem deste trabalho. Inicialmente, apresenta-se uma síntese dos trabalhos existentes, destacando seus objetivos, metodologias empregadas e limitações encontradas.

2.4.1 Trabalhos acadêmicos

As pesquisas acadêmicas relacionadas à modelagem e à análise de dados financeiros concentram-se, majoritariamente, na automatização da análise fundamentalista e na adaptação de modelos tradicionais às especificidades do mercado brasileiro. Nesse contexto, Montoia (2021) desenvolveram um sistema automatizado voltado à análise de ações, com ênfase na celeridade e na eficiência do processo decisório. O estudo descreve a aplicação de algoritmos de processamento de dados que reduzem significativamente a necessidade de intervenção manual, viabilizando a análise de indicadores financeiros em tempo real .

Complementarmente, Guilherme e Marotti (2021) propuseram um modelo adaptável que ajusta os conceitos da análise fundamentalista às particularidades dos dados econômicos nacionais, incorporando variáveis específicas do mercado brasileiro. O referido trabalho ressalta a relevância da personalização dos parâmetros analíticos e da incorporação de variáveis macroeconômicas na modelagem financeira.

No que se refere à melhoria de metodologias consagradas, Delalibera (2023) realizaram aprimoramentos no Modelo Rojo, tradicionalmente utilizado na avaliação de ativos, com ênfase na acurácia da mensuração dos riscos e das oportunidades de investimento. A proposta apresentada integra técnicas estatísticas com algoritmos de aprendizado de máquina (*machine learning*), resultando em ganhos substanciais na precisão das previsões. Por outro lado, Reis (2020) analisaram a aplicação prática desses métodos à realidade do mercado de capitais, evidenciando aspectos relacionados à volatilidade dos ativos e à importância dos diversos indicadores financeiros.

Estudos adicionais, como o de Vieira (2019), abordaram a construção de

carteiras de ações, realizando comparações entre seus desempenhos e índices de referência do mercado. O trabalho discute, ainda, os desafios inerentes à diversificação e ao balanceamento dos ativos. De forma análoga, Freitas (2020) investigaram a utilização de indicadores financeiros no setor bancário, destacando as dificuldades envolvidas na mensuração da solidez financeira e da eficiência operacional das instituições.

2.4.2 Projetos e plataformas similares

No âmbito das iniciativas aplicadas, diversos projetos disponíveis em repositórios digitais, especialmente na plataforma GitHub, têm se destacado por fornecer ferramentas e bases de dados voltadas à extração e análise de informações contábeis. O projeto de PAIVA (2025), por exemplo, apresenta um sistema integrado de coleta, processamento e organização de dados contábeis de empresas, com o objetivo de facilitar a aplicação da análise fundamentalista por meio da disponibilização estruturada e acessível dessas informações.

Adicionalmente, MINAS (2025) propuseram um modelo inovador para a análise de balanços patrimoniais, priorizando a normalização dos dados e a elaboração de indicadores personalizados, ajustados às particularidades de diferentes setores econômicos. Repositórios desenvolvidos por FONTINELE (2025) e LOUREDO (2025) também merecem destaque por viabilizarem a interpretação automatizada de demonstrações financeiras públicas, permitindo a identificação de fragilidades e de oportunidades com maior eficiência.

No entanto, observa-se que grande parte dessas iniciativas concentra-se exclusivamente na coleta e no pré-processamento dos dados, não abrangendo a análise histórica contínua necessária à compreensão das tendências de longo prazo no mercado financeiro.

2.4.3 Diferencial do trabalho proposto

O principal diferencial do presente trabalho consiste na integração de um conjunto histórico abrangente de dados financeiros com um sistema automatizado de acesso e análise das informações públicas disponibilizadas pela Comissão de Valores Mobiliários (CVM). Apesar da delimitação temporal inicial de cinco anos, a solução proposta não está restrita a esse intervalo. A permanência da estrutura atual dos repositórios da CVM assegura a funcionalidade contínua do sistema e a sua capacidade de atualização automática mediante a disponibilização de novos dados.

Destaca-se, ainda, o caráter de código aberto da ferramenta, o que possibilita sua auditoria, personalização e extensão por parte das comunidades acadêmica

e profissional. Essa abertura estimula a colaboração entre pesquisadores e desenvolvedores, promovendo o aperfeiçoamento contínuo do sistema e sua adaptação a diferentes contextos analíticos e setores econômicos.

A metodologia adotada possibilita tanto a avaliação contínua quanto a análise contextualizada das informações. A primeira é viabilizada pelo acúmulo de séries históricas de dados, permitindo a identificação de padrões e tendências ao longo do tempo. A segunda decorre da integração de dados passados e atuais, ampliando a capacidade analítica sobre ciclos econômicos, sazonalidades e eventos que impactam os ativos financeiros.

As inovações e contribuições do sistema desenvolvido podem ser sintetizadas nos seguintes aspectos:

- integração de dados financeiros históricos em larga escala;
- processamento automatizado com ênfase em escalabilidade e reutilização;
- contextualização de indicadores contábeis e financeiros em séries temporais;
- continuidade operacional do sistema frente a eventuais atualizações da CVM, desde que mantida a estrutura original dos dados;
- disponibilização em código aberto, promovendo transparência e incentivo à colaboração técnica e científica.

Ao superar as limitações observadas em iniciativas anteriores, que se restringem majoritariamente à estruturação pontual de dados, a presente proposta constitui uma abordagem sistemática, extensível e fundamentada. Tais características tornam o sistema aplicável a diferentes fins acadêmicos, institucionais e profissionais que demandam análises fundamentadas em princípios econômicos e financeiros.

3 METODOLOGIA

Esta seção apresenta a metodologia de execução do presente trabalho. A Seção 3.1 define o escopo e os objetivos do trabalho, delineando as principais áreas de investigação. Em seguida, a Seção 3.2 descreve as abordagens e técnicas utilizadas para planejar e conduzir o desenvolvimento da ferramenta proposta. Por fim, a Seção 3.3 detalha os passos executados, incluindo a investigação da estrutura dos dados, a modelagem lógica e o desenvolvimento da ferramenta de integração de dados.

3.1 Classificação da pesquisa

Para fundamentar a classificação metodológica da pesquisa, consultaram-se as obras de Cervo e Bervian (1983), que apresentaram diferentes abordagens científicas, e de Gerhardt e Silveira (2009), que trataram de métodos amplamente utilizados na pesquisa acadêmica.

Quanto à natureza, classifica-se a pesquisa como aplicada, pois buscou resolver um problema prático relacionado à integração automatizada de dados financeiros públicos disponibilizados pela CVM. Em relação aos objetivos, a investigação foi exploratória e descritiva. A parte exploratória relacionou-se à compreensão da estrutura dos dados da CVM, enquanto o aspecto descritivo manifestou-se na proposição e implementação de uma ferramenta que permitisse o uso prático dessas informações de forma organizada.

Os procedimentos técnicos combinaram pesquisa bibliográfica e documental para levantamento e compreensão teórica dos dados públicos, juntamente com o desenvolvimento de uma solução tecnológica que automatizasse sua coleta, integração e disponibilização.

A abordagem metodológica adotada foi mista. Aspectos qualitativos estiveram presentes na análise estrutural e na modelagem dos dados, enquanto os elementos quantitativos refletiram-se no processamento automatizado. No contexto da área de Computação, a pesquisa representa o desenvolvimento de uma solução tecnológica voltada à resolução de um problema específico, com potencial de aplicação prática no meio acadêmico e no ambiente de análise financeira (WAZLAWICK, 2009).

3.2 Gerenciamento do projeto

O gerenciamento do projeto foi orientado por práticas ágeis de desenvolvimento de *software*, adotando-se a metodologia *Scrum* como principal referência conceitual (SLIGER, 2011). No entanto, sua aplicação ocorreu de forma adaptada, consi-

derando as especificidades do projeto, a limitação de recursos e o contexto acadêmico em que foi desenvolvido.

Foram empregados ciclos iterativos com base no conceito de *sprints*, com duração aproximada de duas semanas, servindo como guia para o planejamento das atividades e o acompanhamento do progresso. Todavia, não houve a implementação rigorosa de todos os rituais formais do *Scrum*, tais como as reuniões de planejamento, revisão e retrospectiva, nem a adoção de papéis específicos, como *Scrum Master* ou *Product Owner* (SUTHERLAND, 2014).

As decisões sobre prioridades e a definição de tarefas ocorreram, predominantemente, por meio de reuniões regulares com o professor orientador. Esses encontros, de caráter mais flexível, funcionaram como momentos de alinhamento nos quais o andamento do projeto era avaliado e novos direcionamentos eram estabelecidos.

Como ferramenta de apoio à organização do trabalho, foi utilizado um quadro inspirado no modelo *Kanban*, com o objetivo de facilitar a visualização do fluxo de tarefas e permitir um acompanhamento contínuo das entregas (ANDERSON, 2010). Assim como no caso do *Scrum*, o uso do *Kanban* não seguiu todas as prescrições formais da metodologia, sendo adotado de forma pragmática, com foco na clareza e na gestão eficiente das atividades.

Dessa forma, optou-se por uma abordagem híbrida e flexível, que se beneficiou dos princípios fundamentais das metodologias ágeis, sem a necessidade de seguir rigidamente seus modelos estruturados. Tal adaptação mostrou-se adequada às demandas e dinâmicas do ambiente acadêmico, contribuindo para o bom andamento do projeto.

3.3 Solução proposta

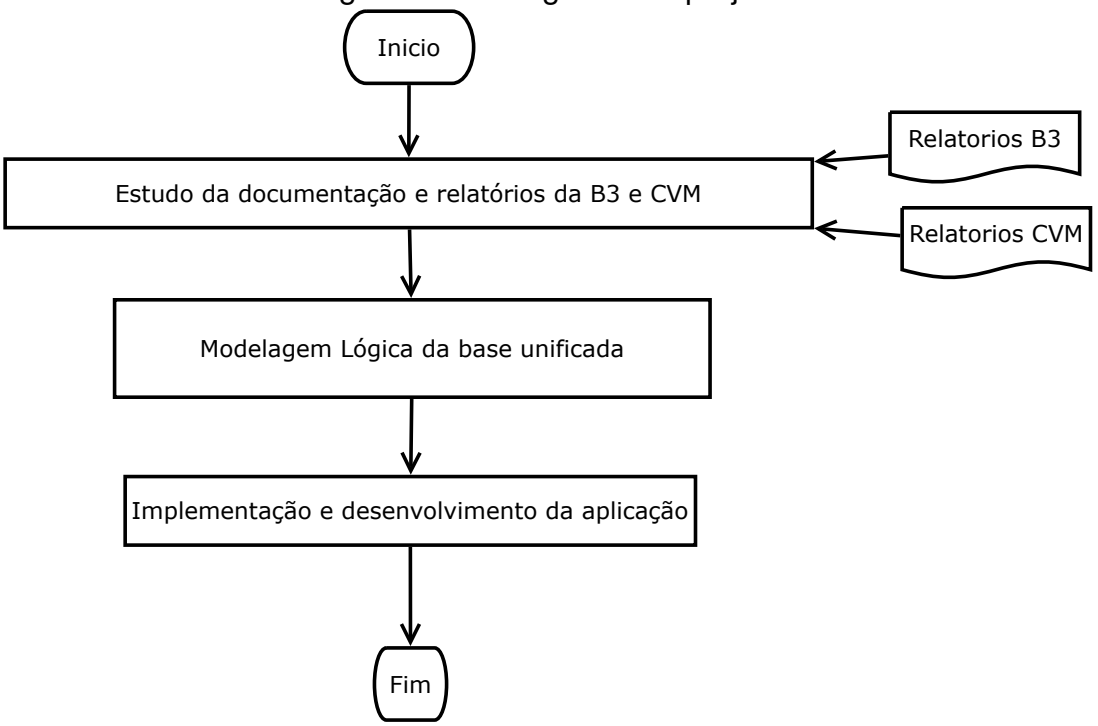
O ponto de partida do projeto consistiu em uma investigação detalhada da estrutura dos dados disponibilizados pela CVM. Essa etapa contemplou a análise dos formatos de arquivos, dos tipos de dados disponíveis e dos métodos de acesso às informações públicas fornecidas por essa instituição.

Embora os dados da B3 não tenham sido utilizados diretamente na construção da ferramenta, sua presença no contexto do mercado de capitais justificou a compreensão de sua estrutura e funcionamento.

Dessa forma, consultaram-se documentos como o relatório anual da CVM (CVM, 2023b) e referências sobre a estrutura de dados da B3 (B3, 2023a), a fim de mapear desafios e possibilidades associados à integração de dados financeiros.

Fundamentado na análise inicial, realizou-se a modelagem lógica da base de dados, com foco na criação de um esquema unificado que contemplasse as dife-

Figura 2 – Fluxograma do projeto



Fonte: Elaborado pelo autor, 2025.

rentes estruturas contábeis e financeiras presentes nos arquivos da CVM.

Para essa etapa, consideraram-se boas práticas descritas na literatura sobre modelagem de dados financeiros (DOMINGUES; ALMEIDA, 2020) e integração de grandes volumes de dados (PERLIN, 2021).

A etapa seguinte consistiu no desenvolvimento da aplicação de *software* responsável por automatizar o processo de coleta, tratamento e inserção dos dados na base unificada. Essa ferramenta foi desenvolvida com tecnologias modernas de extração e processamento de dados, garantindo flexibilidade e desempenho.

A aplicação foi capaz de realizar o *download* dos arquivos disponibilizados pela CVM, aplicar os tratamentos necessários para padronização e limpeza dos dados e armazená-los de forma estruturada.

Esse processo assegurou a integridade das informações e facilitou seu uso em análises fundamentalistas, estudos acadêmicos e outras aplicações. A proposta manteve alinhamento com práticas consolidadas no desenvolvimento de soluções para automação e análise de dados (BESSA; ARTHAUD, 2018).

3.4 Materiais e tecnologias

Esta seção apresenta os materiais e tecnologias empregados durante o desenvolvimento do trabalho. O projeto foi desenvolvido em um computador *desktop* cujas especificações estão descritas no Quadro 1.

Quadro 1 – Especificações do computador utilizado

Componente	**Especificação**
Processador	AMD Ryzen 5 5600G, 6 núcleos, 12 threads, 3.6GHz (até 4.6GHz em turbo), cache de 19MB, soquete AM4
Memória RAM	32GB (2x16GB), DDR4, 3200MHz
Armazenamento	SSD 240GB, SATA III, leitura de até 500 MB/s, gravação de até 450 MB/s HD 1TB, 3.5", 5400 RPM, SATA III, cache de 64MB
Sistema Operacional	Windows 11 Pro

Fonte: Elaborado pelo autor, 2025.

Durante a modelagem do banco de dados, utilizou-se o MySQL Workbench[1], versão 8.0. Essa ferramenta permitiu a criação visual do esquema lógico, facilitando o planejamento e a organização das entidades e relacionamentos.

Contudo, para o desenvolvimento da aplicação, optou-se pelo uso do SQLite[2], versão 3.49.1, como solução de banco de dados local. O SQLite foi adotado como banco de dados padrão do sistema por oferecer maior praticidade na integração com o código desenvolvido.

A linguagem de programação utilizada foi Python, na versão 3.12.9[3], pela sua versatilidade, ampla comunidade e bibliotecas especializadas.

As bibliotecas utilizadas no projeto foram:

- Pandas, versão 2.2.1[4];
- Numpy, versão 1.26.4[5];
- Beautifulsoup4, versão 4.12.2[6];
- Lxml, versão 5.1.0[7];
- Requests, versão 2.31.0[8];
- SQLAlchemy, versão 2.0.27[9];
- Mysql-Connector-Python, versão 8.3.0[10];
- Sqlite3, versão 2.6.0[11].

A biblioteca *Pandas* desempenhou papel central na manipulação e análise de dados tabulares, oferecendo uma estrutura baseada em *dataframes* que se mostrou eficiente para organizar, filtrar e agrupar informações financeiras de maneira

[1] https://www.mysql.com/products/workbench/
[2] https://www.sqlite.org/
[3] https://www.python.org/downloads/release/python-3129/
[4] https://pandas.pydata.org/
[5] https://numpy.org/
[6] https://www.crummy.com/software/BeautifulSoup/
[7] https://lxml.de/
[8] https://requests.readthedocs.io/en/latest/
[9] https://www.sqlalchemy.org/
[10] https://dev.mysql.com/downloads/connector/python/
[11] https://docs.python.org/3/library/sqlite3.html

sistemática.

Complementando o uso do *Pandas*, a biblioteca *NumPy* foi aplicada para realizar operações numéricas vetoriais e matriciais. Essa combinação viabilizou cálculos precisos e o tratamento de grandes volumes de dados com elevada performance computacional.

No processo de extração de informações provenientes de páginas web, empregou-se a biblioteca *BeautifulSoup4*, cuja funcionalidade permite navegar pela estrutura de documentos em *HyperText Markup Language* (HTML, linguagem de marcação de hipertexto) e realizar a extração de dados de forma precisa. Associada a ela, utilizou-se a biblioteca *lxml*, que atua como um analisador sintático (parser) eficiente para documentos HTML e *eXtensible Markup Language* (XML, linguagem de marcação extensível).

A obtenção dos dados disponibilizados pela Comissão de Valores Mobiliários (CVM) foi viabilizada por meio da biblioteca *Requests*, a qual proporcionou comunicação estável com serviços HTTP, possibilitando o *download* automatizado das informações requeridas. A comunicação com o banco de dados foi estruturada com o auxílio do *SQLAlchemy*, ferramenta que forneceu abstração robusta para as operações de inserção, consulta e atualização de dados no sistema SQLite, por meio do paradigma de mapeamento objeto-relacional (ORM).

Durante a etapa inicial de testes com a estrutura relacional modelada em MySQL, recorreu-se à biblioteca *Mysql-Connector-Python*, conector oficial destinado à integração entre aplicações Python e bancos de dados MySQL. Essa integração assegurou plena compatibilidade entre o modelo relacional e os *scripts* desenvolvidos, permitindo a validação adequada da estrutura proposta em um ambiente amplamente consolidado no contexto de bancos de dados relacionais.

Finalizada essa fase de testes, decidiu-se pela migração definitiva para o banco de dados SQLite. A escolha foi motivada por sua leveza, portabilidade e facilidade de integração com sistemas locais, características que se mostraram especialmente vantajosas para estudos exploratórios e reproduzíveis com ênfase acadêmica ou experimental, conforme os objetivos estabelecidos neste trabalho.

4 DESENVOLVIMENTO

Este capítulo apresenta o processo de desenvolvimento da ferramenta proposta para a coleta, estruturação e integração de dados financeiros públicos disponibilizados pela CVM, com foco em análises fundamentalistas de companhias abertas brasileiras. As atividades foram organizadas em três blocos principais, que estruturam as seções deste capítulo.

A Seção 4.1 descreve a análise preliminar dos dados disponibilizados pela CVM, com destaque para a identificação de padrões, inconsistências e limitações nos metadados. A Seção 4.2 detalha a modelagem e estruturação da base de dados, com ênfase na padronização relacional e na aplicação de boas práticas de modelagem para contextos financeiros. Por fim, a Seção 4.3 apresenta o sistema desenvolvido em Python, incluindo os módulos responsáveis pela coleta, transformação e armazenamento dos dados, bem como os mecanismos de registro de eventos.

4.1 Análise inicial dos dados da CVM

Esta seção tem como objetivo explorar as características, formatos e limitações dos dados disponibilizados pela CVM, de forma a selecionar aqueles mais adequados para análises fundamentalistas automatizadas.

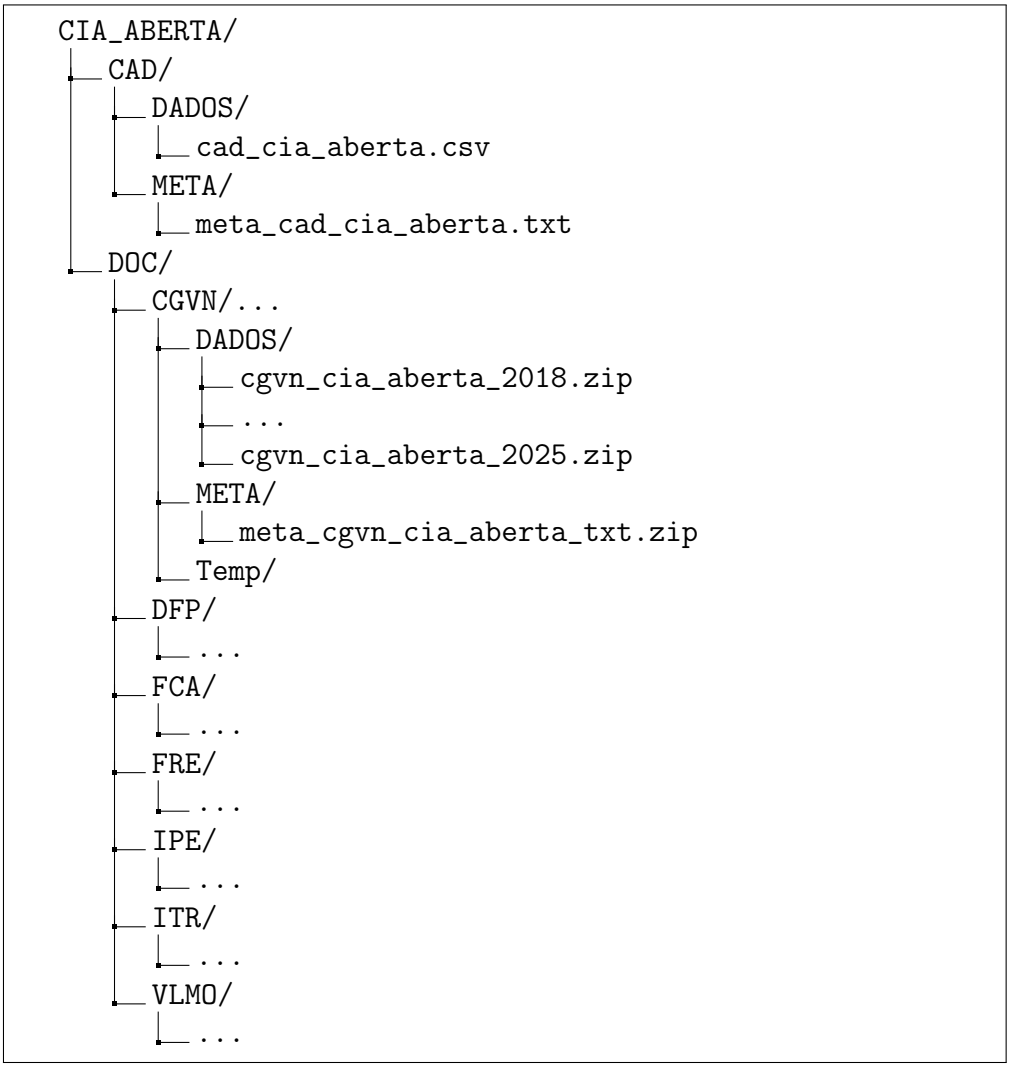
A primeira etapa consistiu na compreensão do formato, da frequência de atualização e de outros aspectos relevantes relacionados aos dados disponibilizados pela CVM nas Companhias (CIA) Abertas. A partir disso, foram identificados os conjuntos de dados, organizados nas seguintes categorias:

- informação cadastral;
- formulário cadastral;
- informações periódicas e eventuais;
- formulário de referência;
- valores mobiliários negociados e detidos;
- formulário de informações trimestrais;
- formulário de demonstrações financeiras padronizadas;
- informe do código de governança.

A partir dessa categorização, a Figura 3 apresenta uma visão geral da estrutura hierárquica dos diretórios e arquivos disponibilizados na seção de Companhias Abertas do portal de dados da CVM. Essa organização é composta por duas pastas principais: `CAD`, que concentra os dados cadastrais das companhias, e `DOC`, que reúne os documentos enviados para CVM.

Dentro da pasta `DOC`, os dados são subdivididos em diretórios específicos conforme o tipo de informação regulatória, como demonstrações financeiras padroni-

Figura 3 – Estrutura hierárquica dos dados das Companhias Abertas no site da CVM



Fonte: Elaborado pelo autor, 2025.

zadas, formulário de informações trimestrais e formulário de referência, entre outros. Cada uma dessas subpastas contém arquivos de dados anuais, disponibilizados no formato compactado (ZIP), além de arquivos de metadados estruturados separadamente na pasta META.

Essa hierarquia padronizada permite a segmentação lógica das informações, facilita o controle de versões e favorece a automação no processo de coleta, extração e análise. Dessa forma, a estrutura adotada pela CVM contribui para a reprodutibilidade dos resultados e para a construção de soluções computacionais robustas que utilizem tais dados como insumo.

As informações cadastrais compõem uma categoria distinta, classificada como cadastro, enquanto os demais arquivos são agrupados sob a categoria de documentos. Ao acessar o site de dados da CVM[1], especificamente na seção referente às Companhias Abertas, é possível visualizar essas duas categorias de dados.

[1] https://dados.cvm.gov.br/dados/

Adicionalmente, é importante contextualizar o sistema responsável pelo envio eletrônico dos documentos à CVM. O sistema *Empresas.NET* (ENET) é a plataforma oficial utilizada pelas companhias abertas para a transmissão de informações regulatórias exigidas pela autarquia. Essa ferramenta garante a padronização, segurança e rastreabilidade do envio, sendo o meio exclusivo de entrega de documentos previstos nas normas vigentes.

Os dados disponibilizados no portal da CVM são apresentados em formatos estruturados, como *comma-separated values* (CSV, valores separados por vírgulas), texto simples no formato `TXT` e arquivos comprimidos no formato `ZIP`. Esses formatos visam garantir maior acessibilidade e automação no tratamento das informações. A maioria dos conjuntos de dados está organizada em arquivos anuais, separados por tipo de formulário e data de referência, o que permite consultas retroativas e reprodutibilidade dos resultados.

Essas pastas listadas anteriormente podem conter dados referentes à categoria principal ou subdivisões internas em (subcategorias), conforme a complexidade do conteúdo. Os metadados associados a cada categoria são apresentados de duas formas. Quando não há subcategorias, os metadados são oferecidos em arquivos de texto. Caso contrário, são fornecidos em arquivos compactados que, ao serem descompactados, revelam arquivos textuais para cada subcategoria.

Os dados propriamente ditos também são disponibilizados em arquivos compactados por ano. Após a extração, o conteúdo desses arquivos pode variar conforme a estrutura da categoria. Em categorias que não possuem subdivisões, o arquivo anual contém um único documento estruturado, abrangendo todas as informações consolidadas. Já em categorias com subcategorias, o mesmo arquivo anual inclui múltiplos documentos estruturados, organizados de acordo com os diferentes tipos de informação tratados.

Essa organização reflete um padrão hierárquico adotado na distribuição dos dados. Pastas correspondentes à categoria principal abrigam os arquivos de metadados e os dados extraídos, com a granularidade ajustada conforme a presença ou ausência de subcategorias.

A disponibilização e estruturação dos dados têm como base legal a Resolução CVM n.º 80/2022 (CVM, 2022), que consolida as regras relativas ao registro e envio de informações pelas companhias abertas. Adicionalmente, a Resolução CVM n.º 44/2021 (CVM, 2021) trata da divulgação de informações relativas a atos ou fatos relevantes, da negociação de valores mobiliários na pendência dessas informações ainda não divulgadas, bem como da divulgação de operações realizadas por pessoas com acesso a informações privilegiadas.

4.1.1 Informação cadastral

O conjunto de informação cadastral reúne os dados cadastrais das companhias abertas disponibilizados pela CVM. Entre as informações fornecidas, destacam-se o número do Cadastro Nacional da Pessoa Jurídica (CNPJ), a data de registro da companhia e a situação atual desse registro. Esses dados são usados para análises regulatórias, econômicas e financeiras, além de servirem como base para estudos acadêmicos e pesquisas de mercado.

O dicionário de dados, disponibilizado em formato textual, contém a descrição detalhada das colunas e dos tipos de dados presentes no arquivo principal. Já os dados propriamente ditos, que contêm os registros cadastrais das companhias abertas, são apresentados em um arquivo estruturado.

Embora o conjunto de informações cadastrais não tenha sido utilizado diretamente nas etapas de extração, modelagem ou análise do presente trabalho, sua estrutura é apresentada a seguir, com o intuito de ilustrar o escopo e a organização dos dados disponibilizados pela CVM.

O Quadro 2 destaca os principais campos presentes nesse conjunto, com suas respectivas descrições e finalidades. Além da amostra apresentada, o Apêndice A apresenta a descrição completa dos campos disponíveis no arquivo principal.

Quadro 2 – Principais campos do conjunto de dados de informação cadastral da CVM

Campo	**Descrição**
`CNPJ_CIA`	Número do Cadastro Nacional da Pessoa Jurídica no formato `NN.NNN.NNN/NNNN-NN`, como `08.773.135/0001-00`. Identificador único para cada companhia, amplamente utilizado como chave primária.
`DENOM_SOCIAL`	Razão social da companhia aberta.
`DT_REG`	Data de registro na CVM, no formato `DD/MM/AAAA`, como `29/10/2020`. Utilizada em análises temporais.
`SIT`	Situação cadastral da companhia, como `ATIVO`, `CANCELADA` ou `EM LIQUIDAÇÃO`. Relevante para segmentações por status operacional.
`CD_CVM`	Código identificador exclusivo atribuído pela CVM, como `25224`. Utilizado na recuperação de documentos e integração com sistemas regulatórios.
`SETOR_ATIV`	Setor de atuação da companhia.
`EMAIL`	Endereço eletrônico institucional da companhia, como `ri@2wecobank.com.br`.

Fonte: Elaborado pelo autor, 2025.

4.1.2 Formulário cadastral

O formulário cadastral (FCA) é um documento eletrônico cuja entrega, seja periódica ou eventual, é regulamentada pela CVM e por suas resoluções. Trata-se de uma obrigação regulatória voltada às companhias abertas, que devem enviá-lo por meio do sistema ENET.

A principal função do FCA é atualizar e manter organizadas as informações institucionais e operacionais dessas entidades. Ele assegura que a CVM e o mercado tenham acesso contínuo a dados relevantes sobre as companhias, contribuindo para a transparência e o bom funcionamento do mercado de capitais.

Os dados públicos relacionados ao FCA, disponibilizados pela própria CVM, estão organizados em duas categorias principais. A primeira diz respeito aos endereços de download dos documentos completos submetidos pelas companhias abertas. A segunda abrange os conteúdos estruturados das diversas seções que compõem o formulário.

Seguindo o padrão já apresentado de subcategorias, o FCA possui uma categoria que se refere ao arquivo `fca_cia_aberta.csv`, contendo os links para o *download* dos documentos completos entregues pelas companhias nos últimos cinco anos. A visualização desses documentos requer a utilização do ENET. As demais categorias compreendem os conteúdos integrais do formulário, organizados por ano, em arquivos estruturados compactados, conforme descrito anteriormente.

A estrutura interna dos dados relacionados ao formulário cadastral (FCA) compreende diversos arquivos organizados por temática. Cada um desses arquivos aborda uma seção específica do formulário, o que permite a segmentação das informações e uma análise mais detalhada dos dados declarados pelas companhias abertas.

No desenvolvimento da ferramenta proposta neste trabalho, foi utilizado exclusivamente o arquivo `fca_cia_aberta_geral.csv`, por concentrar, de forma consolidada, as principais informações institucionais das companhias abertas, dispensando a necessidade de integração dos demais arquivos individualizados. Os principais campos presentes nesse conjunto de dados estão descritos no Quadro 3.

Para fins de análise temporal e organização, os formulários referentes aos anos de 2020 a 2025 encontram-se agrupados em arquivos anuais distintos. Este conjunto de dados integra a base denominada *Documentos Periódicos e Eventuais de Regulados*. Sua manutenção ocorre de forma semanal.

Quadro 3 – Campos presentes no arquivo `fca_cia_aberta_geral.csv`

Campo	Descrição
`CNPJ`	Número do Cadastro Nacional da Pessoa Jurídica.
`Nome Empresarial`	Razão social atual da companhia aberta, armazenada como texto.
`Nome Anterior`	Denominações anteriores da empresa, em formato textual.
`Data Constituição`	Data de fundação da empresa, no formato `DD/MM/AAAA`.
`Setor`	Setor econômico principal em que a companhia atua.
`Descrição da Atividade`	Resumo textual das operações da companhia.
`Situação`	Situação da empresa. Permite identificar o estágio atual de atividade da companhia.
`Website`	Endereço eletrônico oficial da companhia.

Fonte: Elaborado pelo autor, 2025.

4.1.3 Informações periódicas e eventuais

O conjunto de dados referente às informações periódicas e eventuais (IPE) reúne documentos não estruturados enviados por companhias abertas à CVM. Esses documentos, de caráter regulatório, contemplam tanto as obrigações periódicas quanto as comunicações eventuais exigidas ao longo das atividades societárias e da interlocução com o mercado. Trata-se de um acervo que reflete a diversidade de exigências legais e normativas aplicáveis às companhias reguladas. O conteúdo do conjunto está organizado em seis grandes categorias documentais:

- governança e estrutura societária;
- relação com investidores e mercado;
- informações econômico-financeiras e contábeis;
- transações e operações societárias;
- companhias em situação especial;
- informações regulatórias específicas.

Embora o escopo completo do conjunto inclua diversas categorias, este trabalho utiliza apenas parte das informações disponibilizadas, com ênfase nos registros mais diretamente relacionados à estrutura societária e à comunicação obrigatória ao mercado.

Os documentos são atualizados semanalmente para os dois anos mais recentes, incorporando tanto novas submissões quanto eventuais reapresentações. O acervo histórico, por sua vez, remonta a 2003. Cada pacote anual é acompanhado por um dicionário de dados textual, no qual se detalham a lógica de indexação e os campos presentes nos arquivos.

A análise dos registros permite identificar alterações institucionais relevantes, como mudanças de nome empresarial ou de composição acionária. Tais dados

servem como subsídios importantes para o monitoramento de eventos corporativos associados a um determinado *ticker*.

Quadro 4 – Campos presentes nos documentos eventuais submetidos via ENET

Campo	**Descrição**
`CNPJ_Companhia`	Número do Cadastro Nacional da Pessoa Jurídica.
`Nome_Companhia`	Razão social da empresa responsável.
`Código_CVM`	Identificador numérico.
`Tipo`	Categoria do documento submetido.
`Assunto`	Resumo do conteúdo central do documento.
`Tipo_Apresentação`	Modalidade de submissão do documento. Distingue versões iniciais de documentos e suas reapresentações.

Fonte: Elaborado pelo autor, 2025.

Os principais campos disponibilizados no conjunto IPE estão descritos no Quadro 4. Essa estrutura de dados é a organização cronológica e temática dos registros. A presença desses campos possibilita a realização de análises sobre eventos societários, alterações relevantes, reapresentações documentais e outros movimentos estratégicos divulgados ao mercado.

Uma visão geral das diferentes classificações de documentos incluídos no conjunto IPE, com exemplos representativos de documentos e suas respectivas finalidades, é apresentada no Apêndice B.

4.1.4 Formulário de referência

O formulário de referência (FRE) é um documento eletrônico cuja apresentação à CVM é obrigatória e periódica ou, em determinadas circunstâncias, eventual, por meio do sistema ENET. Sua principal finalidade é consolidar e divulgar, de maneira padronizada, um conjunto de informações sobre o emissor, contemplando sua estrutura societária, situação financeira, práticas de governança, riscos e relação com o mercado.

O conjunto de dados públicos associados ao FRE é composto por dois elementos principais:

- os endereços para *download* dos formulários submetidos;
- o conteúdo estruturado extraído desses documentos.

O primeiro item refere-se aos *links* diretos para os formulários completos entregues pelas companhias abertas nos últimos cinco anos, disponibilizados por meio do arquivo `fre_cia_aberta.csv`. Já o segundo item corresponde aos dados tabulares derivados do conteúdo dos formulários, organizados em arquivos estruturados que abordam diversas dimensões informacionais, agrupadas nas seguintes categorias:

- informações institucionais e operacionais;
- aspectos financeiros e patrimoniais;
- administração e governança;
- ações e mercado;
- aspectos sociais e de diversidade.

Essas categorias representam diferentes seções do FRE e permitem análises segmentadas conforme a natureza da informação declarada. As informações estruturadas são atualizadas semanalmente, incorporando tanto novas submissões quanto reapresentações, e mantêm um histórico contínuo desde 2010. Essas atualizações derivam diretamente da base de Documentos Periódicos e Eventuais de Regulados submetidos à CVM.

Cada pacote anual, disponível para os anos de 2010 a 2025, é disponibilizado em formato compactado e contém um conjunto completo de arquivos organizados por tipo de informação, um dicionário de dados, também compactado, que detalha as variáveis e colunas presentes.

Dentre os arquivos incluídos nesses pacotes, destaca-se o `fre_cia_aberta.csv`, que possui especial relevância nesta pesquisa. Esse arquivo concentra, de forma consolidada, as principais informações estruturadas de cada submissão do FRE e inclui os identificadores e os links necessários para recuperação dos documentos completos no sistema ENET.

Embora os demais arquivos disponíveis nos pacotes ofereçam informações relevantes e detalhadas sobre aspectos específicos do formulário, sua estrutura fragmentada e granular, aliada à sobreposição parcial de informações com o arquivo principal, levaram à opção por sua não integração nesta etapa do trabalho. Assim, adotou-se exclusivamente o `fre_cia_aberta.csv` como base de dados para análise, considerando sua abrangência e consistência com os objetivos da pesquisa.

Essa delimitação visa garantir maior simplicidade no modelo de integração dos dados, sem prejuízo da possibilidade de incorporação futura de arquivos complementares, conforme a evolução das necessidades analíticas.

4.1.5 Valores mobiliários negociados e detidos

O conjunto de dados Valores Mobiliários Negociados e Detidos (VLMO) contempla informações de envio obrigatório à CVM de natureza periódica, cujo cumprimento é exigido das companhias abertas. O objetivo principal desse conjunto é registrar e divulgar a posição e as negociações realizadas com valores mobiliários por administradores, membros do conselho fiscal, controladores e pessoas a eles vinculadas. Tais informações são fundamentais para garantir a integridade e a confiança nas práticas de governança corporativa.

Os dados são disponibilizados em arquivos anuais compactados, referentes aos últimos cinco anos, contendo os informes estruturados entregues pelas companhias. Esses arquivos reúnem informações como:

- nome e CPF/CNPJ dos declarantes;
- tipos e quantidades de valores mobiliários detidos;
- natureza da operação (compra, venda, bonificação, entre outros);
- data e características das transações realizadas;
- relação do declarante com a companhia emissora.

Os arquivos são acompanhados de um dicionário de dados, também em formato compactado. O conjunto é atualizado semanalmente. Embora a CVM informe que o histórico disponível abrange os anos de 2020 a 2025, é possível identificar dados desde 2018.

Importa esclarecer que, para este estudo, não foi utilizado nenhum arquivo do conjunto VLMO. Destaca-se, ainda, que esse conjunto, assim como os demais analisados, apresenta uma estrutura padronizada, composta por dois arquivos principais.

O primeiro arquivo, nomeado como *vlmo_cia_aberta.csv*, contém os endereços para download dos documentos entregues pelas companhias. O segundo é um arquivo estruturado que apresenta a consolidação dos dados do VLMO. Ambos os arquivos encontram-se organizados por ano e são disponibilizados em formato compactado.

4.1.6 Demonstrativos financeiros padronizados

Os demonstrativos financeiros padronizados (DFP) são documentos eletrônicos obrigatórios que reúnem informações contábeis anuais estruturadas. De maneira complementar, as informações trimestrais (ITR) também representam documentos eletrônicos de entrega compulsória, porém, com periodicidade trimestral, permitindo um acompanhamento intermediário ao longo do exercício social.

Ambos os relatórios seguem um padrão uniforme de divulgação. Nessas divulgações, estão incluídas as principais demonstrações financeiras, informações cadastrais, pareceres técnicos e arquivos para *download*. As demonstrações financeiras incluídas nesses relatórios são:

- balanço patrimonial ativo (BPA);
- balanço patrimonial passivo (BPP);
- demonstrações dos fluxos de caixa - método direto (DFC-MD);
- demonstrações dos fluxos de caixa - método indireto (DFC-MI);
- demonstração das mutações do patrimônio líquido (DMPL);
- demonstração do resultado (DRE);
- demonstração do resultado abrangente (DRA);

- demonstração do valor adicionado (DVA).

Além das demonstrações financeiras mencionadas, os relatórios também apresentam informações adicionais, tais como dados cadastrais das companhias, pareceres de auditoria independente, declarações emitidas pelos administradores, composição acionária detalhada, estrutura do capital social e links para *download* integral dos documentos submetidos.

Ambos os conjuntos de informações são atualizados semanalmente pela CVM, permitindo análises comparativas e personalizadas, com o auxílio de um dicionário técnico disponibilizado pela própria entidade reguladora, que detalha e explica todas as variáveis e campos contidos nas bases de dados estruturadas.

Apesar de compartilharem um modelo semelhante, esses relatórios se diferenciam essencialmente no período de abrangência. Os dados históricos das DFP remontam ao ano de 2010 e referem-se exclusivamente a exercícios anuais completos, enquanto as ITR encontram-se disponíveis a partir de 2011 e têm como objetivo principal o acompanhamento trimestral.

Quadro 5 – Principais campos presentes no conjunto de dados de DFP

Campo	**Descrição**
`CNPJ`	Número do Cadastro Nacional da Pessoa Jurídica da companhia.
`companhia`	Nome empresarial completo.
`data referencia`	Data de referência da demonstração.
`data final`	Data de encerramento efetivo do exercício ou trimestre.
`versão`	Número inteiro que indica a submissão da demonstração, útil para identificar reapresentações.
`ordem`	Ordenador numérico que define a sequência cronológica entre diferentes versões de um mesmo período.
`CVM`	Código identificador único da companhia atribuído pela CVM.
`grupo`	Tipo e escopo da demonstração contábil.
`moeda`	Unidade monetária da demonstração.
`escala`	Fator de escala aplicado aos valores,.
`código conta`	Identificador numérico da conta contábil, utilizado para padronização entre companhias e períodos.
`descrição conta`	Nome da conta contábil.
`valor conta`	Valor monetário da conta.
`conta fixa`	Indicador de estrutura contábil fixa.

Fonte: Elaborado pelo autor, 2025.

A estrutura dos arquivos dos DFP, exemplificada a partir do BPA, segue um modelo homogêneo de representação de dados contábeis, permitindo comparabilidade entre companhias e ao longo do tempo. Cada linha do arquivo representa o valor de uma conta contábil específica, associada a uma companhia identificada por

seu CNPJ. Além disso, são incluídos metadados fundamentais, como o nome da companhia, a data de referência da demonstração, o código CVM, a moeda e a escala dos valores apresentados.

Os campos versão e ordem possibilitam o rastreio de reapresentações e atualizações dos documentos submetidos, enquanto a composição da estrutura contábil é detalhada por meio de campos como código conta, descrição conta, valor conta e conta fixa. O Quadro 5 resume os principais campos que compõem esse conjunto de dados.

Tais arquivos contemplam versões individuais e consolidadas das companhias, permitindo análises contábeis em diferentes níveis de agregação e comparabilidade. Essa estrutura padronizada permite a leitura automatizada dos demonstrativos e viabiliza comparações intertemporais e entre empresas, com flexibilidade para filtragem por entidade, tipo de demonstração, conta contábil ou período.

4.1.7 Informe do código de governança

O informe do código de governança é um dos instrumentos regulatórios exigidos das companhias abertas brasileiras, com o objetivo de divulgar, de forma estruturada, as práticas de governança corporativa adotadas. Embora seja referenciado no site institucional da CVM como informe do código brasileiro de governança corporativa (ICBGC), nos conjuntos de dados disponibilizados na plataforma de dados abertos, o conteúdo correspondente aparece como código de governança das companhias (CGVN). Considerando essa divergência de nomenclatura, este trabalho adotou a sigla CGVN, conforme empregada nos arquivos estruturados.

O envio do CGVN é de caráter obrigatório e periódico. Esse informe tem como propósito promover a transparência das práticas de governança corporativa. Além disso, permite avaliar o grau de aderência aos princípios fundamentais de governança, como equidade, responsabilidade corporativa, prestação de contas e transparência.

Os principais tópicos abordados no CGVN são:

- estrutura de governança vigente na companhia;
- composição e funcionamento dos órgãos de administração e fiscalização;
- políticas corporativas implementadas;
- nível de aderência às práticas recomendadas pelo CGVN;
- justificativas apresentadas para os casos em que as recomendações não são seguidas.

Os dados abrangem o período de 2018 a 2025 e são atualizados semanalmente, contemplando tanto novas submissões quanto reapresentações. O conjunto de dados CGVN é disponibilizado anualmente em arquivos compactados, organiza-

dos em duas categorias principais: `cgvm_cia_aberta` e `cgvn_cia_aberta_praticas`. O primeiro arquivo contém os endereços para *download* dos documentos completos.

Já o segundo arquivo apresenta os dados estruturados de forma tabular, organizados por companhia e prática declarada. Esse arquivo inclui colunas que indicam o CNPJ da companhia, a data de referência do informe, a versão do documento, o nome empresarial, os identificadores de documento e item, além de campos descritivos, como o capítulo do código, o princípio relacionado, a prática recomendada, a indicação de adoção e, quando aplicável, a justificativa para sua não adoção.

Cabe destacar que, para os fins deste trabalho, o conjunto de dados CGVN não foi incluído na análise prática, uma vez que o foco recai sobre conjuntos mais diretamente relacionados às demonstrações financeiras e ao cadastro das companhias.

4.1.8 Resumo da análise inicial dos dados

Com o intuito de compreender a estrutura e o comportamento dos dados disponibilizados pela CVM, foram desenvolvidos dois scripts em Python, cuja lógica e funcionamento estão descritos nos Apêndices C e D.

O primeiro script é responsável por realizar uma varredura recursiva na pasta pública de dados abertos mantida pela CVM, efetuando o download de todos os arquivos disponíveis. Essa abordagem garante uma coleta completa do repositório público, permitindo posterior análise exploratória do conteúdo obtido.

O segundo script executa a extração em lote dos arquivos compactados, permitindo não apenas o acesso ao conteúdo interno, mas também a mensuração do volume, da diversidade e da granularidade dos dados presentes em cada conjunto. Essa etapa foi essencial para a análise preliminar, fornecendo uma visão concreta da organização e da complexidade das bases tratadas.

Com base na análise exploratória viabilizada pelos scripts desenvolvidos, foi possível identificar quais conjuntos de dados da CVM apresentam maior relevância, cobertura e estrutura adequada aos objetivos deste trabalho. A seleção final concentrou-se em bases que oferecem informações estruturadas, atualizadas e historicamente abrangentes, com potencial para análises fundamentalistas automatizadas.

O Quadro 6 resume os principais aspectos dos conjuntos de dados escolhidos, os quais constituem o núcleo informacional utilizado ao longo deste estudo.

A análise exploratória dos arquivos forneceu informações valiosas para a etapa seguinte, o mapeamento dos dados. Essa fase consistiu na identificação das tabelas e colunas de origem, na avaliação de sua frequência de atualização e na estruturação de um modelo preliminar de destino. As informações coletadas por meio dos metadados auxiliaram na definição de domínios, tipos, tamanhos e descrições de campos, estabelecendo as bases técnicas para a modelagem relacional.

Quadro 6 – Comparativo entre conjuntos de dados da CVM

Aspecto	**ITR**	**DFP**	**FRE**	**FCA**	**IPE**
Tipo de conteúdo	Informações Trimestrais	Demonstrações Financeiras Anuais	Formulário de Referência	Cadastro de Companhia Aberta	Documentos Periódicos / Eventuais
Frequência	Trimestral	Anual	Periódico / Eventual	Periódico / Eventual	Periódico / Eventual
Formato	CSV em ZIP	CSV em ZIP	CSV em ZIP	CSV em ZIP	CSV em ZIP
Volume compactado	366 MB	154 MB	122 MB	5,93 MB	26,3 MB
Volume descompactado	9,62 GB	3,49 GB	941 MB	28,2 MB	261 MB
Estrutura	Demonstrativos contábeis	Demonstrativos consolidados	Informações qualitativas diversas	Dados cadastrais padronizados	Documentos PDF + metadados
Atualização	Semanal	Semanal	Semanal	Semanal	Semanal
Período histórico	Desde 2011	Desde 2010	Desde 2010	Desde 2010	Desde 2003
Utilização	Análise de desempenho trimestral	Análise patrimonial e histórica	Análise estratégica e institucional	Base de entidades	Monitoramento de eventos corporativos

Fonte: Elaborado pelo autor, 2025.

4.2 Modelagem e estruturação dos dados

A modelagem da base de dados teve início com o mapeamento das principais categorias disponibilizadas pela CVM, considerando também, quando existentes, suas respectivas subcategorias. Essa etapa preliminar teve como foco a análise da estrutura dos arquivos e dos metadados associados, permitindo compreender a organização, a granularidade e os padrões recorrentes nos dados fornecidos. O detalhamento completo encontra-se no Apêndice E.

O procedimento adotado seguiu uma lógica semelhante à de um processo *Extract, Transform, Load* (ETL, extração, transformação e carga), com ênfase na padronização e organização das informações extraídas das demonstrações financeiras reportadas por companhias abertas à CVM.

Durante a etapa de extração, foram utilizados arquivos textuais contendo os metadados, disponibilizados em pacotes compactados organizados por tipo de demonstração contábil, como BPA, BPP, DRE, DFC-MD, DFC-MI, DMPL, DRA, DVA e pareceres de auditoria. Cada arquivo trata exclusivamente de um tipo de demonstração.

A etapa de transformação consistiu na padronização dos nomes de colunas e na normalização das estruturas tabulares. Realizou-se o mapeamento entre os campos originais e seus equivalentes no modelo relacional, adotando nomenclatura clara no padrão `snake_case`. Exemplos incluem a substituição de `CD_CVM` por `codigo_cvm` e `DS_CONTA` por `descricao_conta`.

Adicionalmente, foi realizado o enriquecimento temporal dos dados por meio da criação de três campos derivados a partir da data de referência (`DT_REFER`): `data_doc`, `mes_doc` e `ano_doc`. A tipagem dos campos foi definida com base em seus domínios originais, utilizando tipos como `varchar`, `decimal`, `date`, `smallint` e `char`, com destaque para colunas binárias com valores padronizados, como *S* para sim ou *N* para não.

O carregamento dos dados estruturados dos demonstrativos financeiros, independentemente dos tipos de relatórios, foi realizado na tabela `Dfp`, permitindo o agrupamento das informações em uma única estrutura responsável pelo armazenamento dos demonstrativos financeiros. Casos específicos, como os pareceres de auditoria, foram alocados separadamente na tabela `Dfp_Parecer`.

Adicionalmente, foi criada a tabela auxiliar `grupo_dfp`, destinada a indicar se os dados referem-se a demonstrativos individuais ou consolidados. A estrutura resultante desse processo é padronizada e favorece a integração em repositórios analíticos, como *data warehouses*, além de viabilizar consultas temporais e comparações entre companhias. Para isso, foram adotadas estratégias específicas de modelagem, como:

- redução de campos e tabelas desnecessárias;
- definição clara e consistente dos relacionamentos;
- unificação dos demonstrativos contábeis em tabelas únicas;
- criação de índices específicos para acelerar consultas analíticas.

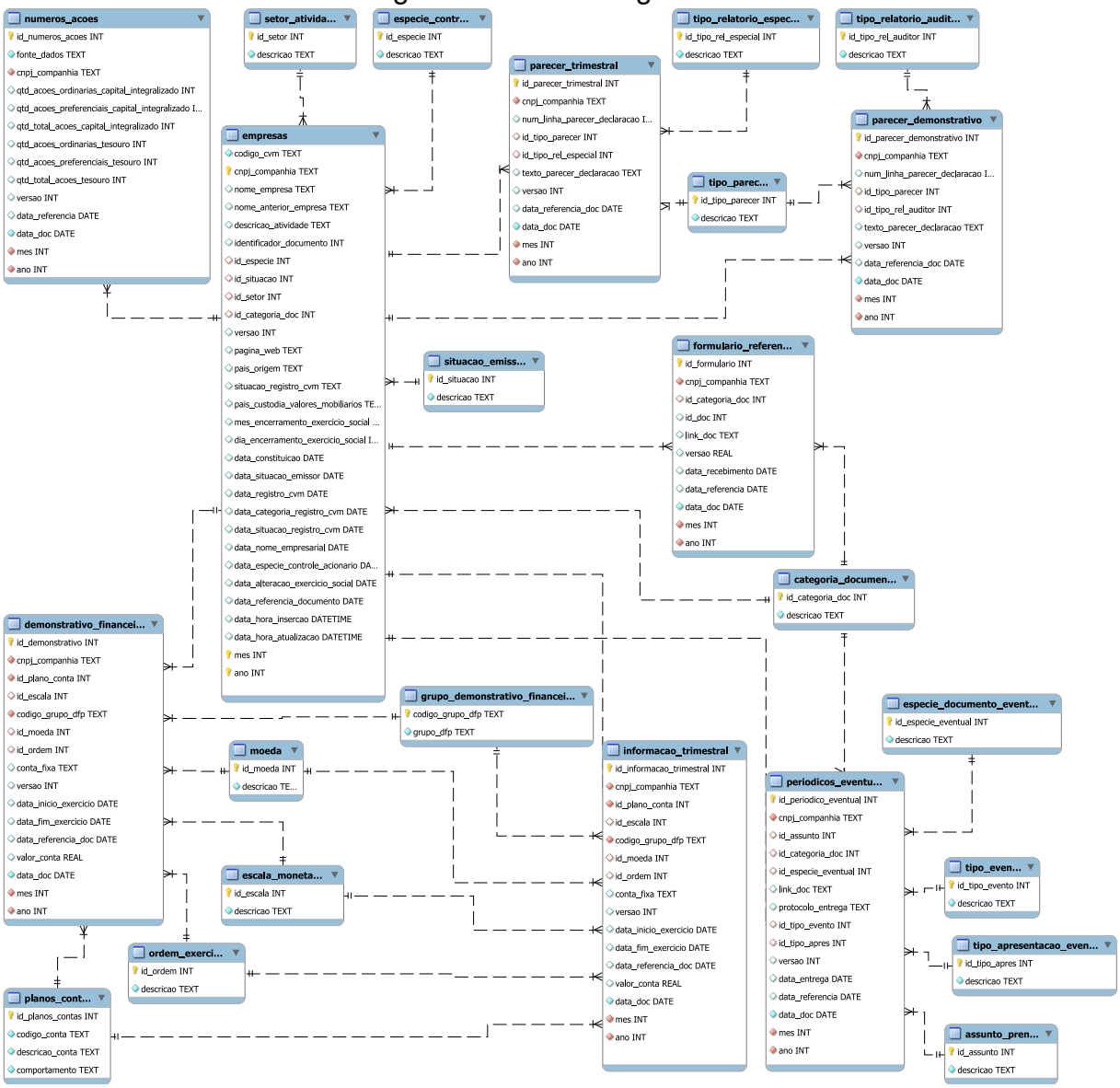
Com base em reuniões periódicas com o orientador, definiu-se que o foco da modelagem seria a otimização da estrutura de dados com vistas à análise fundamentalista, priorizando desempenho, integridade e reprodutibilidade. A Figura 4 apresenta a versão final do modelo lógico, do sistema proposto.

Observa-se a centralidade da tabela `empresas`, com chave primária representada pelo campo `cnpj_companhia`, o que viabiliza a ligação com as demais tabelas. A modelagem favorece análises agregadas por período, setor ou entidade.

Embora a ênfase recaia sobre a versão final do modelo de dados, que corresponde ao modelo já refinado após o processo de normalização, nas formas 2FN e 3FN, é importante destacar que o processo preliminar de refinamento foi essencial para se alcançar uma estrutura adequada. A versão inicial refere-se ao primeiro esboço da base de dados elaborado, servindo como ponto de partida para testes de estruturação lógica. Nesse estágio inicial, o modelo ainda apresentava redundâncias e ausência de padronização, deficiências corrigidas posteriormente, com a aplicação das regras de normalização.

Na sequência, ajustes sucessivos foram conduzidos com o objetivo de eliminar complexidades desnecessárias e promover uma maior simplificação estrutural. A versão refinada da modelagem, com parte dessas melhorias já incorporada, pode

Figura 4 – Modelo lógico final



Fonte: Elaborado pelo autor, 2025.

ser visualizada no Apêndice F. Já o Apêndice G apresenta a penúltima versão, anterior à definitiva, que evidencia a consolidação das principais estratégias adotadas no processo de modelagem.

A estrutura final possui apenas os dados essenciais, com nomenclaturas padronizadas e organização relacional intuitiva. Sua modularidade facilita a integração com outras fontes e ferramentas de análise. Destaca-se, ainda, por ser aberta e pública, podendo ser utilizada em pesquisas acadêmicas, ferramentas de apoio à decisão financeira, além de iniciativas de transparência e educação financeira.

Esses aspectos permitem análises fundamentalistas, especialmente por meio da aplicação de indicadores de rentabilidade, como LPA e P/L, além dos indicadores de liquidez, endividamento e das estimativas consistentes do valor intrínseco das empresas.

Por fim, a padronização dos dados representa um avanço significativo na organização das informações contábeis e corporativas. Essa uniformização permite comparações históricas, oferecendo suporte a diferentes perfis de usuários, como investidores, pesquisadores e analistas financeiros.

No processo de modelagem do banco de dados, optou-se por utilizar o campo `cnpj_companhia` como chave estrangeira principal nas tabelas associadas à entidade `empresas`, em substituição ao código CVM. Tal decisão foi motivada por critérios técnicos e operacionais, visando à maior eficiência e padronização das ligações entre os dados.

O CNPJ é um identificador único, presente em todos os documentos disponibilizados pela CVM, como demonstrações financeiras (DFP e ITR), eventos corporativos, formulários de referência e cadastros. Esse atributo elimina a necessidade de transformações adicionais ou processos de correspondência entre identificadores, otimizando o carregamento e a manutenção da base de dados.

Outro aspecto relevante é que o CNPJ configura-se como uma chave natural e estável, amplamente reconhecida por sistemas externos e instituições como a B3. Isso favorece a integração com outras fontes de dados. Por fim, é importante considerar que o código CVM pode ser alterado ao longo do tempo, especialmente em situações de encerramento ou reabertura de registros de companhias. Nesse sentido, o uso do CNPJ contribui para uma maior estabilidade, coerência e longevidade do modelo, tornando-o mais robusto e preparado para futuras evoluções e integrações com outras bases.

4.3 Software

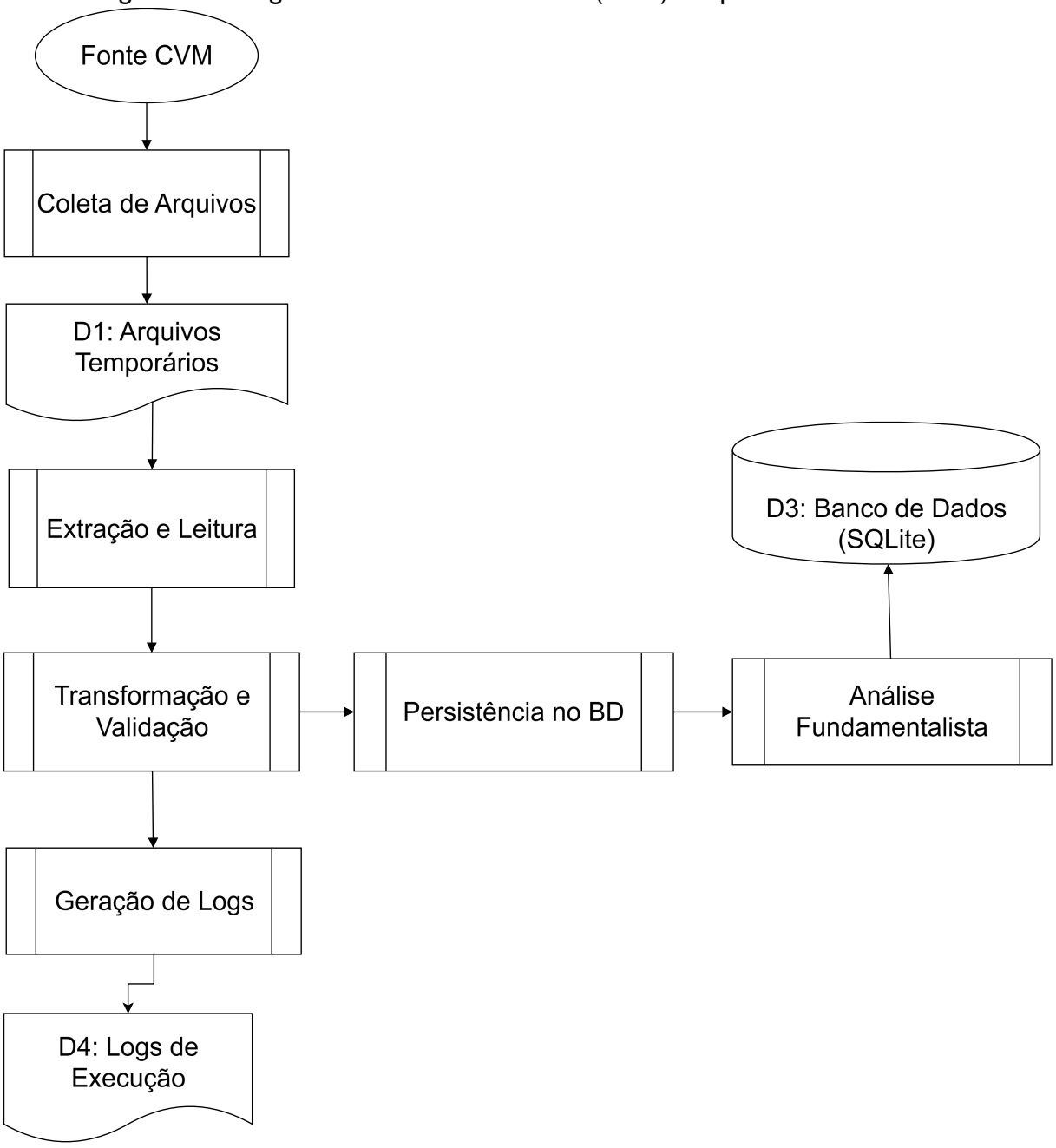
O desenvolvimento do sistema foi realizado de forma incremental, com ênfase na automação das etapas de coleta, extração, transformação e armazenamento dos dados. A arquitetura adotada seguiu um modelo modular, o que favoreceu a organização do código-fonte e contribuiu para a manutenção e evolução do sistema ao longo do tempo.

Com o objetivo de proporcionar uma visão geral de seu funcionamento, a Figura 5 apresenta um Diagrama de Fluxo de Dados (DFD) que ilustra os principais módulos e a interação entre suas respectivas funções, considerando a abordagem de programação estruturada adotada neste trabalho.

O sistema é composto pelos seguintes módulos:

- *Coleta de Arquivos*;
- *Extração e Leitura*;
- *Transformação e Validação*;
- *Persistência no Banco de Dados*;

Figura 5 – Diagrama de Fluxo de Dados (DFD) simplificado do sistema



Fonte: Elaborado pelo autor, 2025.

- *Análise Fundamentalista*;
- *Geração de Logs*.

O módulo *Coleta de Arquivos* é responsável por acessar a estrutura pública da CVM e realizar o *download* automatizado dos arquivos disponibilizados. Na sequência, o módulo *Extração e Leitura* efetua a descompactação dos arquivos obtidos e a leitura de seus conteúdos, preparando-os para os módulos subsequentes.

O módulo *Transformação e Validação* aplica as regras de negócio definidas, realiza os filtros necessários e converte os dados extraídos. Posteriormente, os dados transformados são encaminhados ao módulo *Persistência no Banco de Dados*, no

qual são armazenados, de maneira estruturada, em um banco de dados relacional. No módulo *Análise Fundamentalista*, os dados já persistidos servem de base para o cálculo de indicadores e métricas financeiras.

Todos os módulos do sistema são monitorados pelo módulo *Geração de Logs*, que registra informações relativas a sucessos, falhas e demais eventos relevantes. Esses registros permitem a rastreabilidade do processo e facilitam a identificação de eventuais inconsistências durante a execução.

Por fim, a operação completa do sistema é coordenada por um arquivo principal, que orquestra todo o *fluxo de integração de dados* de processamento. O fluxo geral está descrito na Figura 6.

Figura 6 – Pseudocódigo do fluxo principal de execução

```
iniciar_sistema();
se não houver conexão com a CVM então
    registrar_erro("Sem conexão");
    retorna
arquivos ← coletar_arquivos();
se arquivos = vazio então
    registrar_erro("Nenhum arquivo encontrado");
    retorna
dados_brutos ← extrair_dados(arquivos);
dados_limpos ← transformar_dados(dados_brutos);
persistir_dados(dados_limpos);
para cada conjunto em dados_limpos faça
    realizar_analise_fundamentalista(conjunto);
registrar_logs();
finalizar();
```

Fonte: Elaborado pelo autor, 2025.

Figura 7 – Pseudocódigo da função de coleta de arquivos

```
coletar_arquivos();
para cada url em lista_de_urls faça
    resposta ← requisitar_arquivo(url);
    se resposta estiver OK então
        salvar_arquivo_em_disco(resposta);
        registrar_sucesso(url);
    senão
        registrar_erro(url);
retorna arquivos_baixados;
```

Fonte: Elaborado pelo autor, 2025.

Figura 8 – Pseudocódigo de leitura e normalização de arquivos CSV

```
1 extrair_dados((lista_arquivos));
2 para cada arquivo em lista_arquivos faça
3     tente dados ← ler_csv(arquivo);
4     normalizar_colunas(dados);
5     captura erro registrar_erro(arquivo);
6 retorna(dados_normalizados)
```

Fonte: Elaborado pelo autor, 2025.

Cada etapa é composta por funções específicas. Por exemplo, a função responsável pela coleta de arquivos está representada na Figura 7. O processo de extração e leitura dos dados realiza a padronização das colunas e o tratamento de exceções, conforme ilustrado na Figura 8. A persistência dos dados em banco utiliza uma camada de abstração para garantir integridade transacional. A lógica aplicada à inserção de entidades é exemplificada na Figura 9.

Figura 9 – Pseudocódigo de inserção de entidades no banco de dados

```
1 inserir_entidade((entidade));
2 tente iniciar_transacao();
3 inserir(entidade);
4 confirmar_transacao();
5 captura erro_integridade cancelar_transacao();
6 registrar_alerta(entidade);
```

Fonte: Elaborado pelo autor, 2025.

Todos os eventos do sistema são registrados de forma estruturada em arquivos de log, conforme a Figura 10, que apresenta o padrão de organização dos diretórios de log gerados.

Figura 10 – Exemplo de mensagem registrada em caso de erro

```
2025-05-14 11:34:16,965 - ERROR - Erro ao inserir demonstrativo para CNPJ
↪  00.000.000/0001-91, conta 1, erro: ON CONFLICT clause does not match any
↪  PRIMARY KEY or UNIQUE constraint.
2025-06-06 15:46:37,203 - ERROR - Erro ao inserir Moeda: REAL, erro: UNIQUE
↪  constraint failed: moeda.descricao.
```

Fonte: Elaborado pelo autor, 2025.

Ao longo do desenvolvimento, foram enfrentados desafios relacionados à padronização dos dados, integridade das informações e inconsistências nos arquivos fornecidos pela CVM. Esses aspectos exigiram validações adicionais e refatorações frequentes.

Embora o sistema não possua uma suíte formal de testes automatizados,

cada etapa foi verificada manualmente, com base na análise dos *logs*, inspeção direta no banco de dados e execução segmentada dos *scripts*. Etapas críticas, como a transformação e a consistência de chaves primárias, foram alvo de testes recorrentes para garantir a integridade dos registros.

Ao final, obteve-se um sistema robusto, capaz de realizar o carregamento automatizado de grandes volumes de dados com rastreabilidade e controle de erros. A modularização adotada permite a expansão futura, como a adaptação para outros bancos de dados, a inclusão de testes automatizados e a integração com ferramentas de visualização analítica.

4.4 Módulo para análise fundamentalista

Complementando o *fluxo de integração de dados* de transformação de dados, foi desenvolvido um módulo específico para a realização da análise fundamentalista, com base em indicadores amplamente utilizados no mercado financeiro. Esse módulo tem como objetivo gerar análises a partir dos dados financeiros padronizados extraídos dos formulários ITR e DFP.

A lógica de extração e cálculo dos indicadores foi implementada utilizando-se instruções SQL que combinam diferentes fontes dentro da base de dados, incluindo informações do número de ações, lucro líquido, dividendos, ativos e passivos. Os principais indicadores calculados foram:

- LPA;
- liquidez corrente;
- dívida bruta;
- dívida líquida;
- valor intrínseco.

Para a construção desses cálculos, foram utilizadas cláusulas `WITH`, para organização de subconsultas e agregações condicionais por meio de `CASE WHEN`, de forma a selecionar os valores de contas contábeis conforme seus identificadores. Os dados são agrupados por CNPJ, ano e trimestre, e os resultados são armazenados em novas tabelas dedicadas à análise (`analise_itr` e `analise_dfp`), conforme a Figura 11.

O trecho da consulta SQL que realiza parte dos cálculos de forma estruturada encontra-se apresentado integralmente no Apêndice H. Apesar da robustez do modelo analítico, algumas limitações foram observadas na aplicação prática dos cálculos:

- ausência de dados de mercado;
- dados históricos limitados ou inconsistentes;
- assunções simplificadas.

Figura 11 – Trecho de consulta SQL da análise fundamentalista

```
WITH dados_trimestrais AS (
  SELECT
    itr.cnpj_companhia,
    itr.mes,
    itr.ano,
    (na.qtd_total_acoes_capital_integralizado - na.qtd_total_acoes_tesouro) AS
    ↪  acoes_em_circulacao,
    MAX(CASE WHEN itr.id_plano_conta = '3.13' THEN itr.valor_conta END) AS
    ↪  lucro_liquido,
    MAX(CASE WHEN itr.id_plano_conta = 'Dividendos' THEN itr.valor_conta END)
    ↪  AS dividendos,
    ...
  FROM fato_itr AS itr
  JOIN dim_cadastro_na AS na ON itr.cnpj_companhia = na.cnpj
  ...
)
```

Fonte: Elaborado pelo autor, 2025.

A ausência de dados de mercado é uma limitação importante, uma vez que a base de dados da CVM não fornece diretamente o preço das ações. Isso impede o cálculo direto de indicadores como o P/L e o DY. Para contornar essa limitação, seria necessária a integração com fontes externas, como a B3 ou o Yahoo Finance.

Outra limitação observada refere-se aos dados históricos limitados ou inconsistentes. Nem todas as companhias apresentam séries completas de dados para todos os períodos analisados, o que compromete a análise de tendências ao longo do tempo.

Além disso, o modelo depende de assunções simplificadas, como a definição manual da taxa de desconto e da taxa de crescimento futuro. Essas simplificações reduzem a precisão da estimativa do valor intrínseco, especialmente quando comparadas a modelos mais dinâmicos e adaptados ao contexto de mercado.

Apesar dessas limitações, o sistema ainda representa uma base sólida para análises fundamentalistas automatizadas, com grande potencial de aprimoramento por meio da integração com fontes de dados de mercado e o uso de ferramentas interativas de visualização.

Para demonstrar a aplicabilidade do sistema desenvolvido, elaborou-se uma consulta SQL sobre o banco de dados relacional criado no projeto. Seu objetivo foi extrair os principais indicadores fundamentalistas de companhias do setor financeiro com dados consolidados em dezembro de 2024.

Os campos selecionados abrangeram tanto informações cadastrais das empresas quanto métricas extraídas dos DFP, como LPA, LC, DB e DL. A consulta SQL elaborada para essa finalidade é apresentada na Figura 12.

Figura 12 – Consulta SQL utilizada para gerar os dados da Tabela 1

```sql
SELECT
    e.cnpj_companhia,
    e.nome_empresa,
    e.descricao_atividade,
    a.ano,
    a.mes,
    a.acoes_em_circulacao,
    a.lpa,
    a.lc,
    a.db,
    a.dl
FROM analise_dfp a
INNER JOIN empresas e
    ON e.cnpj_companhia = a.cnpj_companhia
    AND e.ano = a.ano
WHERE a.acoes_em_circulacao IS NOT NULL;
```

Fonte: Elaborado pelo autor, 2025.

A execução da consulta resultou em registros com os principais indicadores fundamentalistas de empresas do setor financeiro. A Tabela 1 apresenta uma amostra desses dados, focada em companhias com informações disponíveis para dezembro de 2024.

Tabela 1 – Resumo dos Indicadores Financeiros das Empresas (dezembro de 2024)

CNPJ	Empresa	Ano	Ações	LPA	LC	DB	DL
30306294000145	BCO BTG PACTUAL S.A.	2024	11.423.711.129	1,03	0,98	502.114.281.000	500.948.264.000
33376989000191	IRB - BRASIL RESSEGUROS S.A.	2024	81.842.887	9,84	1,05	11.520.807.000	11.501.946.000
59285411000113	BCO PAN S.A.	2024	1.250.569.772	0,62	1,10	58.209.318.000	58.203.569.000
60872504000123	ITAU UNIBANCO HOLDING S.A.	2024	9.748.073	3.828,24	1,43	255.592.000.000	250.612.000.000
61186680000174	BANCO BMG S/A	2024	580.482.102	0,72	0,76	46.393.080.000	46.237.308.000
62144175000120	BCO PINE S.A	2024	225.071.879	1,15	1,10	26.616.938.000	26.538.888.000
65654303000173	DIBENS LEASING S.A.	2024	2.215.856	37,24	4,69	72.035.000	71.922.000
91669747000192	DM FINANCEIRA S.A.	2024	61.788	-3.864,52	1,45	2.422.965.000	2.404.468.000

Fonte: Elaborado pelo autor, 2025.

O LPA, por exemplo, apresenta variações expressivas entre empresas, refletindo diferenças operacionais e de desempenho. A LC evidencia a capacidade de pagamento, enquanto os indicadores de endividamento (DB e DL) fornecem uma visão detalhada sobre a estrutura de capital. Esses elementos reforçam a utilidade da base de dados estruturada no suporte a análises financeiras fundamentalistas.

5 CONCLUSÃO

Este trabalho teve como objetivo principal o desenvolvimento de um sistema automatizado capaz de integrar e estruturar os dados financeiros públicos disponibilizados pela CVM, com a finalidade de facilitar o acesso e a aplicação da análise fundamentalista por parte de investidores, analistas e pesquisadores. A motivação para a proposta decorre da constatação de que, embora os dados da CVM estejam amplamente acessíveis, sua forma bruta, descentralizada e tecnicamente fragmentada compromete sua utilização direta em contextos analíticos.

A partir de uma abordagem metodológica mista, composta por pesquisa aplicada, revisão bibliográfica e desenvolvimento prático, foi possível compreender a estrutura dos dados fornecidos pela CVM, realizar a modelagem lógica apropriada e implementar uma ferramenta funcional, desenvolvida na linguagem *Python*, responsável pela coleta, transformação e armazenamento das informações em um banco de dados relacional.

O desenvolvimento do sistema enfrentou desafios técnicos significativos, sobretudo, relacionados à padronização, limpeza e normalização das informações financeiras. A consolidação dos dados em uma base relacional unificada viabilizou consultas consistentes, análises históricas e a construção de indicadores amplamente utilizados na literatura especializada, como o LPA, índices de liquidez, indicadores de endividamento e estimativas de valor intrínseco. Conclui-se que o sistema proposto atingiu os objetivos definidos, oferecendo uma solução viável, acessível e escalável para a análise fundamentalista com base nos dados públicos da CVM.

Tais resultados evidenciam o potencial da ferramenta para servir como suporte a análises financeiras em contextos educacionais, acadêmicos e de mercado. Outro aspecto relevante é o caráter de código aberto adotado no projeto, que favorece a colaboração entre pesquisadores, desenvolvedores e profissionais da área financeira. O código-fonte encontra-se disponível publicamente[1], o que possibilita sua continuidade, personalização e expansão por outros interessados.

Dessa forma, é possível afirmar que as metas inicialmente traçadas foram integralmente alcançadas, conforme sintetizado a seguir:

- análise da estrutura e dos padrões dos conjuntos de dados da CVM, subsidiando as etapas de modelagem e extração;
- projeto e implementação de um modelo relacional, com foco na integridade referencial e normalização;
- desenvolvimento de uma ferramenta automatizada para a coleta, transformação e carga contínua dos dados no banco relacional;
- demonstração, por meio de estudos de caso e consultas `SQL`, da aplicação

[1] Disponível em: https://github.com/ViniciusTAC/cvm-fundamentalist-analysis

prática da base estruturada na construção de indicadores fundamentalistas.

5.1 Limitações

Mesmo com os avanços obtidos, algumas limitações inerentes à natureza dos dados e ao escopo do projeto restringiram parcialmente a abrangência das análises realizadas. A ausência de dados de mercado inviabilizou o cálculo de alguns indicadores, como o P/L e o DY, limitando o escopo da análise fundamentalista. Superar essa limitação exigiria a integração com fontes externas confiáveis, como a B3 ou Yahoo Finance.

A restrição quanto ao número de ações impactou diretamente a possibilidade de se calcular métricas históricas fundamentais, como o LPA. Por se tratar de uma limitação documental da CVM, sua superação dependeria do uso de registros manuais ou alternativos.

Também foram identificadas inconsistências e lacunas nos dados fornecidos pela CVM. Em certos momentos, a documentação indicava a existência de dados a partir de determinado ano, mas os arquivos apresentavam registros de anos anteriores, muitas vezes, sem explicações. Em outros casos, a ausência de documentação sobre a estrutura de algumas tabelas dificultou o correto mapeamento das colunas, exigindo inspeção manual e impactando o tempo de desenvolvimento, a escalabilidade do sistema e a robustez do processo automatizado.

5.2 Trabalhos Futuros

A execução do presente projeto abre caminho para diversas possibilidades de aprimoramento e ampliação da ferramenta desenvolvida. Entre os desdobramentos potenciais, destacam-se:

- expansão da integração para outros dados;
- incorporação de visualizações interativas;
- integração com dados de mercado da B3;
- aplicação de técnicas de aprendizado de máquina.

A base de dados da CVM contempla, além de informações de companhias abertas, dados de fundos de investimento, como fundos imobiliários, multimercado e de renda fixa. A inclusão desses conjuntos ampliaria o escopo analítico do sistema, permitindo investigações mais abrangentes. A implementação de uma interface gráfica conectada ao banco de dados possibilitaria visualizações interativas de indicadores e métricas, favorecendo a acessibilidade do sistema por usuários sem formação técnica.

A integração com bases externas de preços e dados de mercado viabilizaria o cálculo de múltiplos financeiros diretamente relacionados ao valor das ações, conferindo maior completude à plataforma. Com a base estruturada, o sistema pode servir como material para aplicações de aprendizado de máquina, permitindo análises preditivas, classificatórias e exploratórias.

REFERÊNCIAS

ANDERSON, D. J. **Kanban**: Successful Evolutionary Change for Your Technology Business. Washington, DC: Blue Hole Press, 2010.

ATTIE, P. I. **O mercado financeiro e a sustentabilidade**: o papel das bolsas de valores. 2013. Dissertação (Mestrado em Desenvolvimento Econômico) – Universidade Estadual de Campinas (Unicamp), Campinas, SP. Disponível em: https://doi.org/20.500.12733/1621829. Acesso em: 14/03/2024.

BABALOLA, Y. A.; ABIOLA, F. R. Financial ratio analysis of firms: a tool for decision making. **International Journal of Management Sciences**, Academia, San Francisco, v. 1, n. 4, p. 132–137, 2013. Disponível em: https://www.academia.edu/5180572/Financial_Ratio_Analysis_of_Firms_A_Tool_for_Decision_Making. Acesso em: 14/06/2025.

BANCO SANTANDER. **O que é o mercado de capitais?** 2024. Disponível em: https://www.santander.com.br/blog/o-que-e-mercado-de-capitais. Acesso em: 23/04/2024.

BESSA, T.; ARTHAUD, D. D. B. Metodologias ágeis para o desenvolvimento de softwares. **Ciência e Sustentabilidade (CeS)**, Universidade Federal do Cariri (UFCA), Juazeiro do Norte, v. 4, n. 2, p. 173–213, 2018. Disponível em: https://doi.org/10.33809/2447-4606.422018173-213. Acesso em: 07/07/2024.

BRASIL. **Lei n° 6.385, de 7 de dezembro de 1976**, 1976a. Dispõe sobre o mercado de valores mobiliários e cria a Comissão de Valores Mobiliários (CVM). Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l6385.htm. Acesso em: 02/02/2025.

BRASIL. **Lei n° 6.404, de 15 de dezembro de 1976: Dispõe sobre as sociedades por ações**, 1976b. Conhecida como Lei das Sociedades por Ações ou Lei das S.A. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l6404consol.htm. Acesso em: 05/08/2025.

BRASIL, BOLSA, BALCÃO (B3). **Dados de mercado**. 2023a. Disponível em: https://www.b3.com.br/pt_br/market-data-e-indices/servicos-de-dados/market-data/consultas/mercado-a-vista/dados-de-mercado/. Acesso em: 23/04/2024.

BRASIL, BOLSA, BALCÃO (B3). **Número de investidores na B3 cresce 34% em renda fixa e 23% em renda variável em 12 meses**. 2023b. Disponível em: https://www.b3.com.br/pt_br/noticias/numero-de-investidores-na-b3-cresce-34-em-renda-fixa-e-23-em-renda-variavel-em-12-meses.htm. Acesso em: 14/01/2024.

BRASIL, BOLSA, BALCÃO (B3). **Relatório Anual 2023**. 2023c. Disponível em: https://ri.b3.com.br/pt-br/informacoes-financeiras/relatorio-anual. Acesso em: 23/04/2024.

CERVO, A. L.; BERVIAN, P. A. **Metodologia científica**. São Paulo: McGraw-Hill, 1983.

CLAESSENS, S.; KOSE, A. M. Macroeconomic implications of financial imperfections: a survey. **CEPR Discussion Papers**, 2017. Disponível em: https://ssrn.com/sol3/abstract=3076410. Acesso em: 10/03/2024.

CODD, E. F. Relational database: a practical foundation for productivity. **Communications of the ACM**, v. 25, n. 2, p. 109–117, 1982. Disponível em: https://doi.org/10.1145/358396.358400. Acesso em: 14/06/2025.

COMISSÃO DE VALORES MOBILIÁRIOS (CVM). CVM. **Resolução CVM nº 80, de 29 de março de 2022**, 2022. Dispõe sobre o registro e a prestação de informações periódicas e eventuais dos emissores de valores mobiliários admitidos à negociação em mercados regulamentados de valores mobiliários. Disponível em: https://conteudo.cvm.gov.br/legislacao/resolucoes/resol080.html. Acesso em: 14/06/2025.

COMISSÃO DE VALORES MOBILIÁRIOS (CVM). **Funções da CVM**. 2023a. Disponível em: https://www.gov.br/cvm/pt-br/acesso-a-informacao-cvm/institucional/competencia. Acesso em: 06/05/2024.

COMISSÃO DE VALORES MOBILIÁRIOS (CVM). **O que é a CVM?** 2009. Disponível em: https://www.gov.br/cvm/pt-br/acesso-a-informacao-cvm/servidores/estagio/2-materia-cvm-e-o-mercado-de-capitais. Acesso em: 25/02/2024.

COMISSÃO DE VALORES MOBILIÁRIOS (CVM). **Relatório Anual da CVM**. 2023b. Disponível em: https://www.gov.br/cvm/pt-br/centrais-de-conteudo/publicacoes/relatorios/anual. Acesso em: 25/02/2024.

COMISSÃO DE VALORES MOBILIÁRIOS (CVM). **Resolução CVM nº 44, de 23 de agosto de 2021**, 2021. Dispõe sobre a divulgação de informações sobre ato ou fato relevante, a negociação de valores mobiliários na pendência de ato ou fato relevante não divulgado e a divulgação de informações sobre a negociação de valores mobiliários, e revoga as Instruções CVM nº 358/2002, nº 369/2002 e nº 449/2007. Disponível em: https://conteudo.cvm.gov.br/legislacao/resolucoes/resol044.html. Acesso em: 14/06/2025.

CONNOLLY, T.; BEGG, C. **Database Systems**: A Practical Approach to Design, Implementation, and Management. Harlow: Pearson, 2015.

DAMODARAN, A. **Investment Valuation**: Tools and Techniques for Determining the Value of Any Asset. 3. ed. Hoboken: Wiley, 2012.

DANTAS, M. **Comportamento da bolsa de valores no Brasil diante das crises globais de 2008 e 2020**. 2020. Trabalho de Conclusão de Curso (Bacharelado em Ciências Econômicas) – Pontifícia Universidade Católica (PUC) de Goiás, Goiânia. Disponível em: https://repositorio.pucgoias.edu.br/jspui/handle/123456789/1251. Acesso em: 14/01/2024.

DELALIBERA, W. **Automatização do Modelo Rojo de Análise Fundamentalista**. 2023. Dissertação (Mestrado em Administração) – Universidade Estadual do Oeste do Paraná (UNIOESTE), Cascavel. Disponível em: https://tede.unioeste.br/handle/tede/6666. Acesso em: 08/08/2025.

DEREK, B. *et al.* Intrinsic Value: A Solution to the Declining Performance of Value Strategies. **Financial Analysts Journal**, Taylor & Francis, v. 81, n. 2, p. 67–88, 2025. Disponível em: https://doi.org/10.1080/0015198X.2025.2467027. Acesso em: 14/06/2025.

DOMINGUES, O.; ALMEIDA, G. L. M. Modelagem de cointegração em ativos financeiros. In: CONGRESSO BRASILEIRO DE ENGENHARIA DE PRODUÇÃO (CONBREPRO), X., 2020, Evento on-line. **Anais [...]** Curitiba: Universidade Tecnológica Federal do Paraná (UTFPR), 2020. p. 1–11. Disponível em: https://aprepro.org.br/conbrepro/2020/anais/. Acesso em: 10/03/2024.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de banco de dados**. São Paulo: Pearson Addison Wesley, 2005.

FIGUEIREDO, P. N. **Capacidade tecnológica e inovação**: desafios para a transição industrial e econômica do Brasil. Rio de Janeiro: FGV Editora, 2023.

FONTINELE, A. **Análise de Dados**: Balanços de Empresas CVM Dados Públicos. 2025. Disponível em: https://github.com/Alfredo-Fontinele/Analise_Dados_CVM. Acesso em: 11/02/2025.

FREITAS, L. C. d. **Uma análise fundamentalista do setor bancário**: um estudo de caso com uma seleção de indicadores. 2020. Trabalho de Conclusão de Curso (Graduação em Finanças) – Universidade Federal do Ceará (UFC), Fortaleza. Disponível em: http://www.repositorio.ufc.br/handle/riufc/60417. Acesso em: 08/08/2025.

FUNDAÇÃO GETULIO VARGAS (FGV). **Pesquisa da FGV aponta caminhos para transformação econômica do Brasil por meio da inovação**. 2024. Disponível em: https://rededepesquisa.fgv.br/noticia/pesquisa-da-fgv-aponta-caminhos-para-transformacao-economica-do-brasil-por-meio-da. Acesso em: 07/07/2024.

GERHARDT, T. E.; SILVEIRA, D. **Métodos de pesquisa**. Porto Alegre: Editora da UFRGS, 2009.

GÓIS, A. D.; SOARES, S. V. O efeito da reputação corporativa segundo a transparência contábil no gerenciamento de resultados de empresas listadas na B3. **Revista de Educação e Pesquisa em Contabilidade (REPeC)**, v. 13, n. 2, 2019. Disponível em: https://www.repec.org.br/repec/article/view/2229. Acesso em: 25/02/2024.

GOMES, F. R. A Bolsa de Valores brasileira como fonte de informações financeiras. **Perspectivas em Ciência da Informação**, v. 2, n. 2, 2007. Disponível em: https://periodicos.ufmg.br/index.php/pci/article/view/23238/18793. Acesso em: 07/07/2024.

GUILHERME, D. A. G. d. A.; MAROTTI, T. R. Modelo de dados flexível para análise fundamentalista moderna: um estudo geral com dados no Brasil. **FGV RIC Revista de Iniciação Científica**, v. 2, 2021. Disponível em: https://periodicos.fgv.br/ric/article/view/86064/81092. Acesso em: 02/02/2025.

HALEVY, A. Y.; RAJARAMAN, A.; ORDILLE, J. Data Integration: The Teenage Years. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES (VLDB), XXXII., 2006, Seoul, Korea. **Proceedings [...]** VLDB Endowment, ACM, 2006. p. 9–16.

JOBIM, K. V. F. **Análise do valor intrínseco das empresas brasileiras (2019–2023)**: com base na fórmula de Benjamin Graham. 2025. Trabalho de Conclusão de Curso (Graduação em Ciências Contábeis) – Universidade Federal do Rio Grande do Norte (UFRN), Natal. Disponível em: https://repositorio.ufrn.br/handle/123456789/63025. Acesso em: 08/08/2025.

KEARNEY, C.; LIU, S. Textual sentiment in finance: A survey of methods and models. **International Review of Financial Analysis**, v. 33, p. 171–185, 2014. Disponível em: https://doi.org/10.1016/j.irfa.2014.02.006. Acesso em: 14/03/2024.

KOTHARI, S. P. Capital markets research in accounting. **Journal of Accounting and Economics**, v. 31, n. 1, p. 105–231, 2001. Disponível em: https://doi.org/10.1016/S0165-4101(01)00030-1. Acesso em: 22/03/2024.

LIAW, K. T. **The business of investment banking**: A comprehensive overview. Hoboken: John Wiley & Sons, Ltd, 2011.

LOUREDO, G. **Códigos para ETL, análise estatística e visualização de dados da CVM**. 2025. Disponível em: https://github.com/GrlouX/AnaDadosFRECIAberta. Acesso em: 11/02/2025.

MAURICE, O.; AGYARKO, A. K.; PAUL, A. I. Time-frequency analysis of behaviourally classified financial asset markets. **Research in International Business and Finance**, v. 50, p. 54–69, 2019. Disponível em: https://doi.org/10.1016/j.ribaf.2019.04.012. Acesso em: 14/03/2024.

MINAS, T. **Balanco empresas dados CVM**. 2025. Disponível em: https://github.com/thaisminas/Balanco_Empresas_Dados_CVM. Acesso em: 11/02/2025.

MONTOIA, G. R. **Automatização da análise fundamentalista do mercado de ações brasileiro**. 2021. Trabalho de Conclusão de Curso (Graduação em Engenharia de Controle e Automação) – Universidade Federal de Uberlândia (UFU), Uberlândia. Disponível em: https://repositorio.ufu.br/handle/123456789/36480. Acesso em: 02/02/2025.

PAIVA, J. P. R. D. **Projeto Dados CVM**. 2025. Disponível em: https://github.com/joaopedrordepaiva/DadosCVM. Acesso em: 11/02/2025.

PERLIN, M. S. **Análise de Dados Financeiros e Econômicos com o R**. 3. ed. Porto Alegre: Publicação Independente, 2021.

RAHM, E.; DO, H. H. Data cleaning: problems and current approaches. **Bulletin of the Technical Committee on Data Engineering**, v. 23, n. 4, p. 3–13, 2000. Disponível em: https://nbn-resolving.org/urn:nbn:de:bsz:15-qucosa2-329680. Acesso em: 14/06/2025.

REIS, L. V. B. **Análise fundamentalista aplicada às ações negociadas na bolsa de valores brasileira**. 2020. Trabalho de Conclusão de Curso (Graduação em Ciências Contábeis) – Universidade Federal de Uberlândia (UFU), Uberlândia. Disponível em: https://repositorio.ufu.br/handle/123456789/30547. Acesso em: 08/08/2025.

REIS, T. **Negociação de ações nas bolsas de valores**: Tudo sobre o Mercado Financeiro: o que é e como ele funciona? 2021. Disponível em: https://www.suno.com.br/guias/mercado-financeiro/. Acesso em: 07/07/2024.

ROBERTO, A. M. **Papel da informação contábil auditada no processo de tomada de decisão**: uma pesquisa realizada com futuros gestores. 2023. Trabalho de Conclusão de Curso (Graduação em Administração) – Universidade Federal de Ouro Preto (UFOP), Mariana. Disponível em: http://www.monografias.ufop.br/handle/35400000/6082. Acesso em: 25/02/2024.

SAHU, S. K.; MOKHADE, A.; BOKDE, N. D. An overview of machine learning, deep learning, and reinforcement learning-based techniques in quantitative finance: recent progress and challenges. **Applied Sciences**, v. 13, n. 3, 2023. Disponível em: https://doi.org/10.3390/app13031956. Acesso em: 10/03/2024.

SANTOS, P. d. **Transformação digital no sistema bancário**: o impacto dos bancos digitais no mercado financeiro no Brasil. 2023. Trabalho de Conclusão de Curso (Bacharelado em Administração) – Universidade de Passo Fundo (UPF), Passo Fundo, RS. Disponível em: http://repositorio.upf.br/handle/riupf/2542. Acesso em: 08/08/2025.

SLIGER, M. Agile project management with Scrum. In: PMI GLOBAL CONGRESS, XI., 2011, Dallas. **Proceedings [...]** Newtown Square: Project Management Institute, 2011. p. 1–9. Disponível em: https://irantypist.com/media/new_research/samplefile/1621186018_5646.pdf. Acesso em: 08/08/2025.

SOUZA FIGUEIREDO, C. A. de *et al.* O mercado acionário brasileiro: possibilidades para investir no setor. In: ENCONTRO INTERNACIONAL DE GESTÃO, DESENVOLVIMENTO E INOVAÇÃO (EIGEDIN), V., 2021, online. **Anais [...]** Campo Grande: Universidade Federal de Mato Grosso do Sul (UFMS), 2021. Disponível em: https://periodicos.ufms.br/index.php/EIGEDIN/article/view/13996. Acesso em: 07/07/2024.

SUTHERLAND, J. **Scrum**: the art of doing twice the work in half the time. New York, NY: Crown Currency, 2014.

TEIXEIRA, J. C. **Mercado de Capitais**: O que é e como funciona? 2019. Disponível em: https://fia.com.br/blog/mercado-de-capitais/. Acesso em: 23/04/2024.

VIEIRA, R. K. d. M. **Montagem de carteiras de ações listadas na B3 e comparação com índices de investimentos brasileiros sob a ótica de análise fundamentalista**. 2019. Trabalho de Conclusão de Curso (Graduação em Ciências Contábeis) – Universidade Federal da Paraíba (UFPB), João Pessoa. Disponível em: https://repositorio.ufpb.br/jspui/handle/123456789/17188. Acesso em: 08/08/2025.

WAZLAWICK, R. S. **Metodologia de Pesquisa para Ciência da Computação**. Rio de Janeiro: Elsevier, 2009.

APÊNDICES

APÊNDICE A - DICIONÁRIO DE DADOS DAS INFORMAÇÕES CADASTRAL

Campo	Descrição	Tipo	Tamanho
AUDITOR	Nome do Auditor	varchar	100
BAIRRO	Bairro	varchar	100
BAIRRO_RESP	Bairro do responsável	varchar	100
CATEG_REG	Categoria do registro	varchar	20
CD_CVM	Código CVM	numeric	7
CEP	CEP	numeric	8
CEP_RESP	CEP do responsável	numeric	8
CNPJ_AUDITOR	CNPJ do Auditor	varchar	20
CNPJ_CIA	CNPJ da companhia	varchar	20
COMPL	Complemento de endereço	varchar	100
COMPL_RESP	Complemento do responsável	varchar	100
CONTROLE_ACIONARIO	Controle Acionário	varchar	30
DDD_FAX	Código DDD (FAX)	varchar	4
DDD_FAX_RESP	Código DDD (FAX) do responsável	varchar	4
DDD_TEL	Código DDD (Telefone)	varchar	4
DDD_TEL_RESP	Código DDD (Telefone) do responsável	varchar	4
DENOM_COMERC	Denominação Comercial	varchar	100
DENOM_SOCIAL	Denominação Social	varchar	100
DT_CANCEL	Data de cancelamento	date	10
DT_CONST	Data de constituição	date	10
DT_INI_CATEG	Início da categoria do registro	date	10
DT_INI_RESP	Início do responsável	date	10
DT_INI_SIT	Início da situação	date	10
DT_INI_SIT_EMISSOR	Início da situação do emissor	date	10
DT_REG	Data de registro	date	10
EMAIL	E-mail	varchar	100
EMAIL_RESP	E-mail do responsável	varchar	100
FAX	FAX	numeric	15
FAX_RESP	FAX do responsável	numeric	15

Continua...

Campo	Descrição	Tipo	Tamanho
LOGRADOURO	Logradouro	varchar	100
LOGRADOURO_RESP	Logradouro do responsável	varchar	100
MOTIVO_CANCEL	Motivo de cancelamento	varchar	80
MUN	Município	varchar	100
MUN_RESP	Município do responsável	varchar	100
PAIS	País	varchar	100
PAIS_RESP	País do responsável	varchar	100
RESP	Nome do responsável	varchar	100
SETOR_ATIV	Setor de atividade	varchar	100
SIT	Situação	varchar	40
SIT_EMISSOR	Situação do emissor	char	80
TEL	Telefone	numeric	15
TEL_RESP	Telefone do responsável	numeric	15
TP_ENDER	Tipo de endereço	char	30
TP_MERC	Tipo de mercado	varchar	50
TP_RESP	Tipo de responsável	varchar	100
UF	Unidade da Federação	char	2
UF_RESP	UF do responsável	char	2

APÊNDICE B - LISTA DE DOCUMENTOS DISPONÍVEIS NO CONJUNTO IPE

Documento	Descrição	Categoria Geral
Acordo de Acionistas	Regras pactuadas entre sócios sobre controle e voto	Governança
Assembleia	Atas e deliberações de assembleias gerais	Governança
Aviso aos Acionistas	Informações operacionais e societárias relevantes	Investidores
Aviso aos Debenturistas	Comunicados voltados aos detentores de debêntures	Investidores
Calendário de Eventos Corporativos	Agenda de eventos obrigatórios da companhia	Investidores
Carta Anual de Governança Corporativa	Declaração sobre práticas de governança	Governança
Código de Conduta	Normas internas de comportamento e ética	Governança
Comunicado ao Mercado	Divulgações públicas não obrigatórias	Investidores
Fato Relevante	Divulgação obrigatória de eventos materiais	Investidores
Informações sobre Falência, Liquidação ou Recuperação	Status jurídico-financeiro da companhia	Situação Especial
Plano de Remuneração Baseado em Ações	Programas de incentivos aos executivos	Transações
Política de Divulgação	Normas sobre divulgação de informações sensíveis	Regulatória
Política de Sustentabilidade	Diretrizes para práticas sustentáveis	Governança
Regimento Interno do Conselho de Administração	Regras de funcionamento do conselho	Governança

Continua...

Documento	Descrição	Categoria Geral
Edital de Oferta Pública de Ações (OPA)	Oferta para aquisição de ações no mercado	Transações
Escritura de Debêntures	Termos das emissões de debêntures	Econômico-financeira
Documentos de Oferta Pública	Informações para investidores em IPOs e follow-ons	Econômico-financeira
Informações às Bolsas Estrangeiras	Relatórios encaminhados a bolsas internacionais	Investidores
Política de Dividendos	Critérios para distribuição de lucros	Econômico-financeira
Outras Políticas Regulatórias	Conjunto de normas exigidas pela CVM	Regulatória

APÊNDICE C - CÓDIGO PARA BAIXAR OS DADOS DA CVM

```python
import requests
from bs4 import BeautifulSoup
import os
import time


# Função para fazer o download do arquivo
def download_file(url, save_path):
    response = requests.get(url)
    with open(save_path, "wb") as f:
        f.write(response.content)


# Função para fazer uma requisição HTTP com retries
def make_request(url, retries=5, delay=5):
    for i in range(retries):
        try:
            response = requests.get(url)
            response.raise_for_status()
            return response
        except requests.exceptions.RequestException as e:
            print(f"Erro ao acessar {url}: {e}")
            if i < retries - 1:
                print(f"Tentando novamente em {delay} segundos...")
                time.sleep(delay)
            else:
                raise


# URL base inicial
base_url = "https://dados.cvm.gov.br/dados/CIA_ABERTA/"

# Diretório base de destino
dest_dir_base = "CIA_ABERTA"

# Lista para armazenar os diretórios a serem explorados, iniciando com o
↪ URL base
dirs_to_explore = [base_url]

```

```
39 # Loop para explorar os diretórios
40 while dirs_to_explore:
41     # Remove o primeiro diretório da lista
42     current_dir = dirs_to_explore.pop(0)
43 
44     # Faz a requisição HTTP para obter o conteúdo da página
45     response = make_request(current_dir)
46     soup = BeautifulSoup(response.text, "html.parser")
47 
48     # Encontra todos os links na página
49     for link in soup.find_all("a"):
50         href = link.get("href")
51 
52         # Ignora os links que são de volta ('../')
53         if href == "../":
54             continue
55 
56         # Constrói a URL completa do link
57         full_url = current_dir + href
58 
59         # Se o link for um diretório, adiciona à lista de diretórios a
           ↪ serem explorados
60         if href.endswith("/"):
61             dirs_to_explore.append(full_url)
62         else:
63             # Se o link for um arquivo, verifica se é um arquivo CSV, ZIP
               ↪ ou TXT e faz o download
64             if href.endswith(".csv") or href.endswith(".zip") or
               ↪ href.endswith(".txt"):
65                 # Obtém o caminho relativo do arquivo em relação à URL
                   ↪ base
66                 relative_path = os.path.relpath(full_url, base_url)
67                 # Constrói o caminho completo de destino
68                 save_dir = os.path.join(dest_dir_base,
                   ↪ os.path.dirname(relative_path))
69                 # Cria os diretórios necessários se não existirem
70                 os.makedirs(save_dir, exist_ok=True)
71                 # Constrói o caminho completo do arquivo de destino
72                 save_path = os.path.join(save_dir,
                   ↪ os.path.basename(full_url))
73                 # Faz o download do arquivo
```

```
74                download_file(full_url, save_path)
75                print("Arquivo baixado:", save_path)
76
77    # Imprime o diretório atual
78    print("Diretório atual:", current_dir)
```

APÊNDICE D - CÓDIGO PARA EXTRAIR OS DADOS DA CVM

```python
import os
import zipfile

# Caminhos de origem e destino
pasta_zip = r"CIA_ABERTA"
pasta_destino = os.path.join(os.path.dirname(pasta_zip), "CIA_ABERTA_extraida")

# Percorrer a pasta de origem
for root, _, files in os.walk(pasta_zip):
    for file in files:
        if file.endswith(".zip"):
            caminho_zip = os.path.join(root, file)

            # Caminho relativo da subpasta em relação à raiz da pasta_zip
            caminho_relativo = os.path.relpath(root, pasta_zip)

            # Criar o caminho correspondente na pasta de destino
            destino_subpasta = os.path.join(pasta_destino, caminho_relativo)
            os.makedirs(destino_subpasta, exist_ok=True)

            # Extrair o conteúdo do ZIP nesse caminho
            try:
                with zipfile.ZipFile(caminho_zip, 'r') as zip_ref:
                    zip_ref.extractall(destino_subpasta)
                print(f"Extraído: {file} para {destino_subpasta}")
            except zipfile.BadZipFile:
                print(f"Erro: Arquivo ZIP inválido - {caminho_zip}")
```

APÊNDICE E - MAPEAMENTO COMPLETO DOS DADOS DA CVM

E.1 Dados Cadastrais (DFP)

Arquivo de origem: `meta_dfp_cia_aberta.txt`.

Col. Origem	Col. Dest.	Descrição	Tipo	Domínio	Tamanho	Tab. Dest.
CATEG_DOC	categoria_doc	Categoria do documento	varchar	Alfanumérico	20	Dfp
CD_CVM	codigo_cvm	Código CVM	char	Numérico	6	Dfp
CNPJ_CIA	cnpj_companhia	CNPJ da companhia	varchar	Alfanumérico	20	Dfp
DENOM_CIA	denominacao_companhia	Nome empresarial da companhia	varchar	Alfanumérico	100	Dfp
DT_RECEB	data_recebimento_doc	Data da recebimento do documento	date	AAAA-MM-DD	10	Dfp
DT_REFER	data_referencia_doc	Data de referência do documento	date	AAAA-MM-DD	10	Dfp
ID_DOC	id_doc	Identificador do documento	int	Numérico	10	Dfp
LINK_DOC	link_doc	Endereço para download do documento	varchar	Alfanumérico	121	Dfp
VERSAO	versao	Versão do documento	smallint	Numérico	5	Dfp
data_doc	data_doc	Data do documento (aspecto temporal)	date	AAAA-MM-DD	10	Dfp
mes_doc	mes_doc	Mês do documento (aspecto temporal)	varchar	Numérico	4	Dfp
ano_doc	ano_doc	Ano do documento (aspecto temporal)	varchar	Numérico	4	Dfp

E.2 Balanço Patrimonial Ativo (BPA)

Arquivo de origem: `meta_dfp_cia_aberta_BPA.txt`.

Col. Origem	Col. Dest.	Descrição	Tipo	Domínio	Tamanho	Tab. Dest.
CD_CONTA	codigo_conta	Código da conta	varchar	Numérico	18	Dfp
CD_CVM	codigo_cvm	Código CVM	char	Numérico	6	Dfp
CNPJ_CIA	cnpj_companhia	CNPJ da companhia	varchar	Alfanumérico	20	Dfp
DENOM_CIA	denominacao_companhia	Nome empresarial da companhia	varchar	Alfanumérico	100	Dfp
DS_CONTA	descricao_conta	Descrição da conta	varchar	Alfanumérico	100	Dfp
DT_FIM_EXERC	data_fim_exercicio	Data fim do exercício social	date	AAAA-MM-DD	10	Dfp
DT_REFER	data_referencia_doc	Data de referência do documento	date	AAAA-MM-DD	10	Dfp
ESCALA_MOEDA	escala_monetaria	Escala monetária	varchar	Alfanumérico	100	Dfp
GRUPO_DFP	grupo_dfp	Nome e nível de agregação da demonstração	varchar	Alfanumérico	206	grupo_dfp
MOEDA	moeda	Moeda	varchar	Alfanumérico	100	Dfp
ORDEM_EXERC	ordem_exercicio	Ordem do exercício social	varchar	Alfanumérico	9	Dfp
ST_CONTA_FIXA	tipo_conta	Indica se é conta fixa ou não	varchar	S/N	1	Dfp
VERSAO	versao	Versão do documento	smallint	Numérico	5	Dfp
VL_CONTA	valor_conta	Valor da conta	decimal	Numérico	29	Dfp
data_doc	data_doc	Data do documento (aspecto temporal)	date	AAAA-MM-DD	10	Dfp
mes_doc	mes_doc	Mês do documento (aspecto temporal)	varchar	Numérico	4	Dfp
ano_doc	ano_doc	Ano do documento (aspecto temporal)	varchar	Numérico	4	Dfp

E.3 Balanço Patrimonial Passivo (BPP)

Arquivo de origem: `meta_dfp_cia_aberta_BPP.txt`.

Col. Origem	Col. Dest.	Descrição	Tipo	Domínio	Tamanho	Tab. Dest.
CD_CONTA	codigo_conta	Código da conta	varchar	Numérico	18	Dfp
CD_CVM	codigo_cvm	Código CVM	char	Numérico	6	Dfp
CNPJ_CIA	cnpj_companhia	CNPJ da companhia	varchar	Alfanumérico	20	Dfp
DENOM_CIA	denominacao_companhia	Nome empresarial da companhia	varchar	Alfanumérico	100	Dfp
DS_CONTA	descricao_cotna	Descrição da conta	varchar	Alfanumérico	100	Dfp
DT_FIM_EXERC	data_fim_exercicio	Data fim do exercício social	date	AAAA-MM-DD	10	Dfp
DT_REFER	data_referencia_doc	Data de referência do documento	date	AAAA-MM-DD	10	Dfp
ESCALA_MOEDA	escala_monetaria	Escala monetária	varchar	Alfanumérico	100	Dfp
GRUPO_DFP	grupo_dfp	Nome e nível de agregação da demonstração	varchar	Alfanumérico	206	grupo_dfp
MOEDA	moeda	Moeda	varchar	Alfanumérico	100	Dfp
ORDEM_EXERC	ordem_exercicio	Ordem do exercício social	varchar	Alfanumérico	9	Dfp
ST_CONTA_FIXA	conta_fixa	Indica se é conta fixa ou não	varchar	S/N	1	Dfp
VERSAO	versao	Versão do documento	smallint	Numérico	5	Dfp
VL_CONTA	valor_conta	Valor da conta	decimal	Numérico	29	Dfp
data_doc	data_doc	Data do documento (aspecto temporal)	date	AAAA-MM-DD	10	Dfp
mes_doc	mes_doc	Mês do documento (aspecto temporal)	varchar	Numérico	4	Dfp
ano_doc	ano_doc	Ano do documento (aspecto temporal)	varchar	Numérico	4	Dfp

E.4 Demonstração de Fluxo de Caixa - Método Direto (DFC-MD)

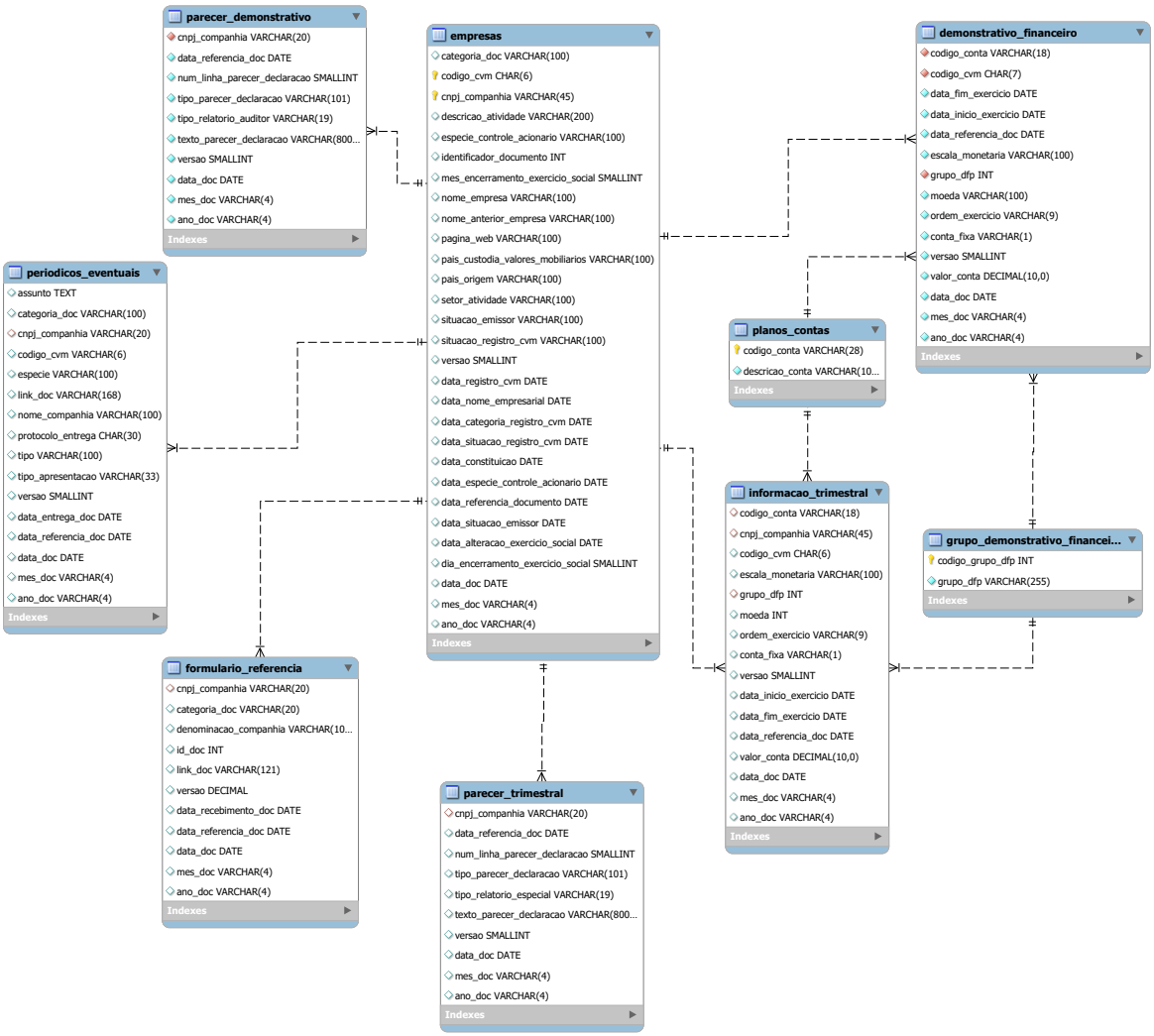
Arquivo de origem: `meta_dfp_cia_aberta_DFC_MD.txt`.

Col. Origem	Col. Dest.	Descrição	Tipo	Domínio	Tamanho	Tab. Dest.
CD_CONTA	codigo_conta	Código da conta	varchar	Numérico	18	Dfp
CD_CVM	codigo_cvm	Código CVM	char	Numérico	6	Dfp
CNPJ_CIA	cnpj_companhia	CNPJ da companhia	varchar	Alfanumérico	20	Dfp
DENOM_CIA	denominacao_companhia	Nome empresarial da companhia	varchar	Alfanumérico	100	Dfp
DS_CONTA	descricao_cotna	Descrição da conta	varchar	Alfanumérico	100	Dfp
DT_FIM_EXERC	data_fim_exercicio	Data fim do exercício social	date	AAAA-MM-DD	10	Dfp
DT_INI_EXERC	data_inicio_exercicio	Data início do exercício social	date	AAAA-MM-DD	10	Dfp
DT_REFER	data_referencia_doc	Data de referência do documento	date	AAAA-MM-DD	10	Dfp
ESCALA_MOEDA	escala_monetaria	Escala monetária	varchar	Alfanumérico	100	Dfp
GRUPO_DFP	grupo_dfp	Nome e nível de agregação da demonstração	varchar	Alfanumérico	206	grupo_dfp
MOEDA	moeda	Moeda	varchar	Alfanumérico	100	Dfp
ORDEM_EXERC	ordem_exercicio	Ordem do exercício social	varchar	Alfanumérico	9	Dfp
ST_CONTA_FIXA	conta_fixa	Indica se é conta fixa ou não	varchar	S/N	1	Dfp
VERSAO	versao	Versão do documento	smallint	Numérico	5	Dfp
VL_CONTA	valor_conta	Valor da conta	decimal	Numérico	29	Dfp
data_doc	data_doc	Data do documento (aspecto temporal)	date	AAAA-MM-DD	10	Dfp
mes_doc	mes_doc	Mês do documento (aspecto temporal)	varchar	Numérico	4	Dfp

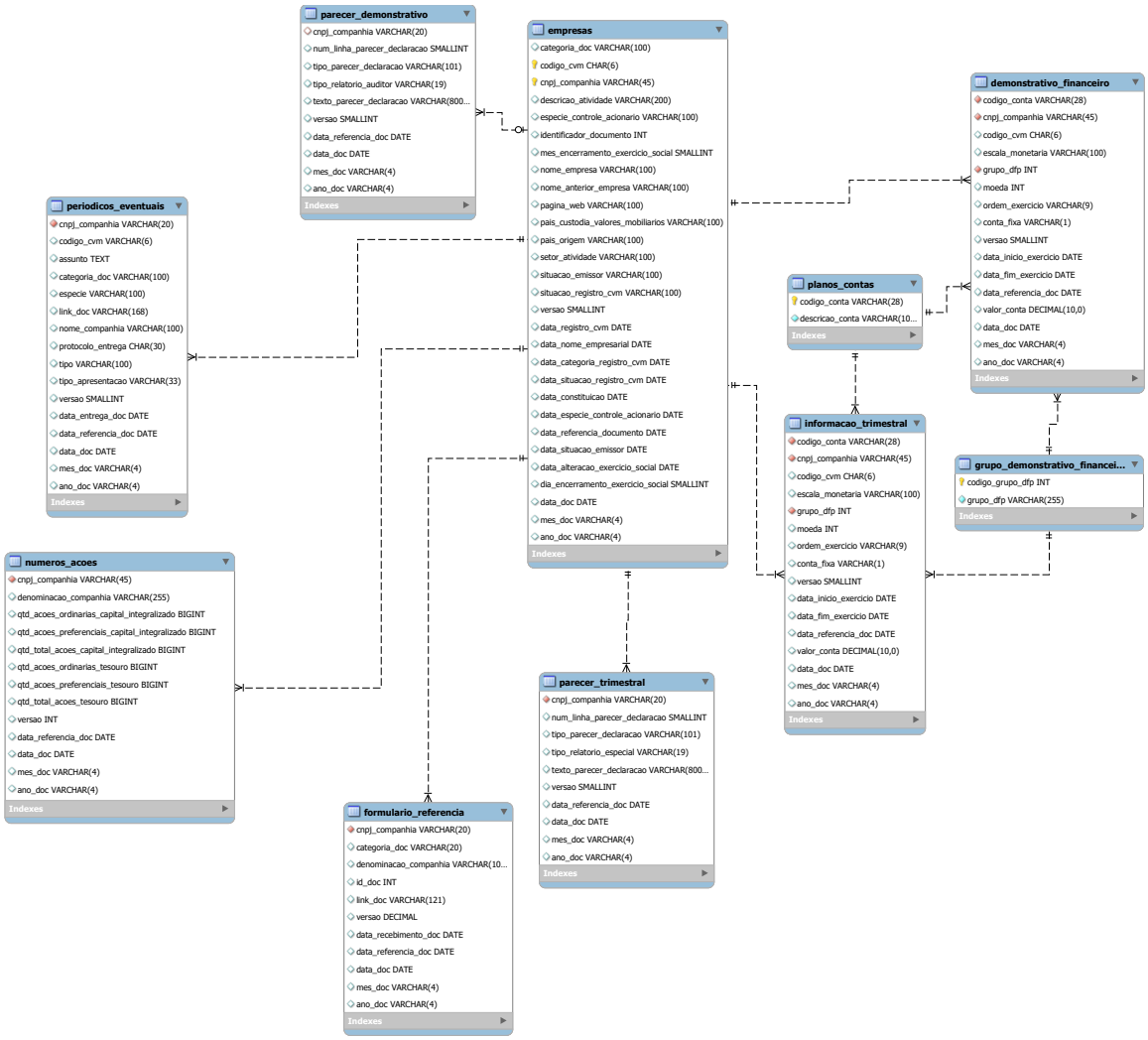
Continua...

Origem	Col. Dest.	Descrição	Tipo	Domínio	Tamanho	Tab. Dest.
ano_doc	ano_doc	Ano do documento (aspecto temporal)	varchar	Numérico	4	Dfp

APÊNDICE F - MODELAGEM APÓS AJUSTES INICIAIS



APÊNDICE G - VERSÃO INTERMEDIÁRIA DO ESQUEMA LÓGICO



APÊNDICE H - CÓDIGO SQL PARA ANÁLISE FUNDAMENTALISTA

```sql
-- ITR
WITH escala_fatores AS (
  SELECT
    id_escala,
    CASE descricao
      WHEN 'MIL' THEN 1e3
      WHEN 'UNIDADE' THEN 1e0
      ELSE 1
    END AS fator
  FROM escala_monetaria
),
dados_trimestrais AS (
  SELECT
    itr.cnpj_companhia,
    itr.mes,
    itr.ano,
    (na.qtd_total_acoes_capital_integralizado -
    ↪ na.qtd_total_acoes_tesouro) AS acoes_em_circulacao,

    MAX(CASE WHEN pc.codigo_conta = '3.13'       THEN itr.valor_conta *
    ↪ COALESCE(ef.fator, 1) END) AS lucro_liquido,
    MAX(CASE WHEN pc.codigo_conta = 'Dividendos' THEN itr.valor_conta *
    ↪ COALESCE(ef.fator, 1) END) AS dividendos,
    MAX(CASE WHEN pc.codigo_conta = '1.01'       THEN itr.valor_conta *
    ↪ COALESCE(ef.fator, 1) END) AS ativo_circulante,
    MAX(CASE WHEN pc.codigo_conta = '2.01'       THEN itr.valor_conta *
    ↪ COALESCE(ef.fator, 1) END) AS passivo_circulante,
    MAX(CASE WHEN pc.codigo_conta = '2.02'       THEN itr.valor_conta *
    ↪ COALESCE(ef.fator, 1) END) AS passivo_nao_circulante,
    MAX(CASE WHEN pc.codigo_conta = '1.01.01'    THEN itr.valor_conta *
    ↪ COALESCE(ef.fator, 1) END) AS disponibilidades

  FROM informacao_trimestral itr
  LEFT JOIN planos_contas pc ON pc.codigo_conta = itr.id_plano_conta
  LEFT JOIN escala_fatores ef ON ef.id_escala = itr.id_escala
  LEFT JOIN numeros_acoes na
    ON na.cnpj_companhia = itr.cnpj_companhia
   AND na.mes = itr.mes
   AND na.ano = itr.ano
```

```
33
34    WHERE pc.codigo_conta IN
      ↪  ('3.13','Dividendos','1.01','2.01','2.02','1.01.01')
35
36    GROUP BY itr.cnpj_companhia, itr.mes, itr.ano,
37             na.qtd_total_acoes_capital_integralizado,
38             na.qtd_total_acoes_tesouro
39  )
40
41  SELECT
42    d.cnpj_companhia,
43    d.ano,
44    d.mes,
45    d.acoes_em_circulacao,
46
47    -- 1) LPA
48    d.lucro_liquido / NULLIF(d.acoes_em_circulacao, 0)                AS lpa,
49
50    -- 2) Liquidez Corrente
51    d.ativo_circulante / NULLIF(d.passivo_circulante, 0)              AS lc,
52
53    -- 3) Dívida Bruta
54    (d.passivo_circulante + d.passivo_nao_circulante)                 AS db,
55
56    -- 4) Dívida Líquida
57    (d.passivo_circulante + d.passivo_nao_circulante)
58      - d.disponibilidades                                            AS dl
59
60  FROM dados_trimestrais d
61  WHERE d.lucro_liquido IS NOT NULL
62  ORDER BY d.ano, d.mes;
63
64  ------------------------------------------------
65
66  -- DFP
67  WITH escala_fatores AS (
68    SELECT
69      id_escala,
70      CASE descricao
71        WHEN 'MIL' THEN 1e3
72        WHEN 'UNIDADE' THEN 1e0
```

```
73        ELSE 1
74      END AS fator
75    FROM escala_monetaria
76  ),
77  dados_anuais AS (
78    SELECT
79      dfp.cnpj_companhia,
80      dfp.mes,
81      dfp.ano,
82      (na.qtd_total_acoes_capital_integralizado -
        ↪  na.qtd_total_acoes_tesouro) AS acoes_em_circulacao,
83
84      MAX(CASE WHEN pc.codigo_conta = '3.13'       THEN dfp.valor_conta *
        ↪  COALESCE(ef.fator, 1) END) AS lucro_liquido,
85      MAX(CASE WHEN pc.codigo_conta = 'Dividendos' THEN dfp.valor_conta *
        ↪  COALESCE(ef.fator, 1) END) AS dividendos,
86      MAX(CASE WHEN pc.codigo_conta = '1.01'       THEN dfp.valor_conta *
        ↪  COALESCE(ef.fator, 1) END) AS ativo_circulante,
87      MAX(CASE WHEN pc.codigo_conta = '2.01'       THEN dfp.valor_conta *
        ↪  COALESCE(ef.fator, 1) END) AS passivo_circulante,
88      MAX(CASE WHEN pc.codigo_conta = '2.02'       THEN dfp.valor_conta *
        ↪  COALESCE(ef.fator, 1) END) AS passivo_nao_circulante,
89      MAX(CASE WHEN pc.codigo_conta = '1.01.01'    THEN dfp.valor_conta *
        ↪  COALESCE(ef.fator, 1) END) AS disponibilidades
90
91    FROM demonstrativo_financeiro dfp
92    LEFT JOIN planos_contas pc ON pc.codigo_conta = dfp.id_plano_conta
93    LEFT JOIN escala_fatores ef ON ef.id_escala = dfp.id_escala
94    LEFT JOIN numeros_acoes na
95      ON na.cnpj_companhia = dfp.cnpj_companhia
96     AND na.mes = dfp.mes
97     AND na.ano = dfp.ano
98
99    WHERE pc.codigo_conta IN
      ↪  ('3.13','Dividendos','1.01','2.01','2.02','1.01.01')
100
101   GROUP BY dfp.cnpj_companhia, dfp.mes, dfp.ano,
102            na.qtd_total_acoes_capital_integralizado,
103            na.qtd_total_acoes_tesouro
104 )
105
```

```sql
SELECT
  d.cnpj_companhia,
  d.ano,
  d.mes,
  d.acoes_em_circulacao,

  -- 1) LPA
  d.lucro_liquido / NULLIF(d.acoes_em_circulacao, 0)              AS lpa,

  -- 2) Liquidez Corrente
  d.ativo_circulante / NULLIF(d.passivo_circulante, 0)            AS lc,

  -- 3) Dívida Bruta
  (d.passivo_circulante + d.passivo_nao_circulante)              AS db,

  -- 4) Dívida Líquida
  (d.passivo_circulante + d.passivo_nao_circulante)
    - d.disponibilidades                                         AS dl

FROM dados_anuais d
WHERE d.lucro_liquido IS NOT NULL
ORDER BY d.ano, d.mes;
```