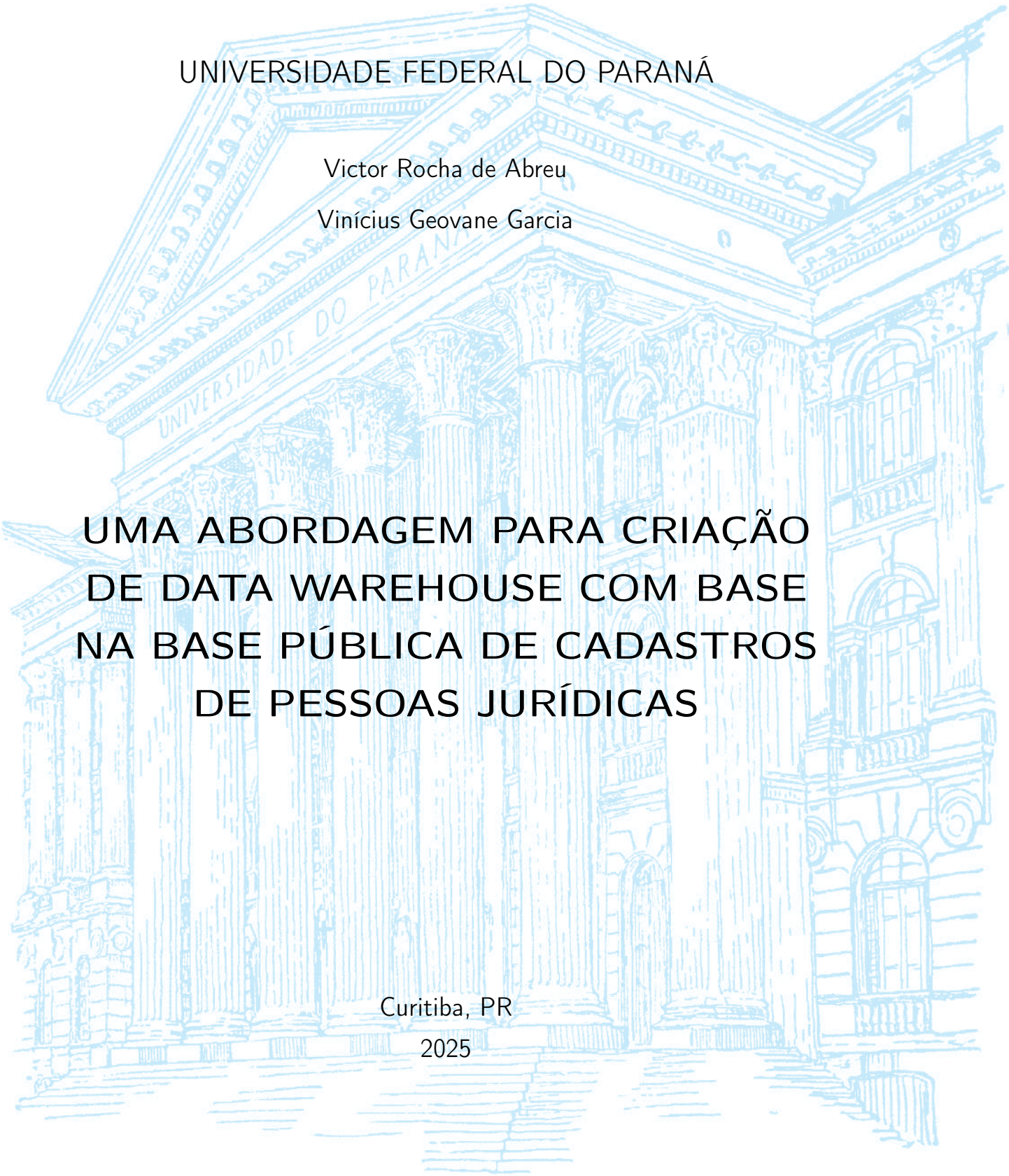
UNIVERSIDADE FEDERAL DO PARANÁ

Victor Rocha de Abreu

Vinícius Geovane Garcia



UMA ABORDAGEM PARA CRIAÇÃO DE DATA WAREHOUSE COM BASE NA BASE PÚBLICA DE CADASTROS DE PESSOAS JURÍDICAS

Curitiba, PR

2025

VICTOR ROCHA ABREU
VINÍCIUS GEOVANE GARCIA

UMA ABORDAGEM PARA CRIAÇÃO DE DATA WAREHOUSE COM BASE NA BASE PÚBLICA DE CADASTROS DE PESSOAS JURÍDICAS

Trabalho de Conclusão de Curso, do Curso de Bacharelado em Ciência da Computação do Departamento de Informática – DINF – da Universidade Federal do Paraná – UFPR, como requisito para a aprovação na disciplina.
Prof. Dr. Andrey Ricardo Pimentel.

CURITIBA
2025

RESUMO

Nos últimos anos, as áreas de ciência e engenharia de dados vêm se consolidando. São importantes e ajudam a extrair informação e conhecimento da grande quantidade de dados que temos disponíveis hoje em dia. A demanda por dados implica em métodos para o armazenamento e estruturação dos dados. Este trabalho tem como objetivo um estudo de caso de uma abordagem para criação de *data warehouses* que preconizem a qualidade, reusabilidade, eficiência e confiabilidade dos dados. Com a intenção de realizar este estudo, entendemos os dados públicos de Cadastro Nacional de Pessoas Jurídicas como uma oportunidade de dados de fácil acesso para utilização da abordagem proposta. Dessa maneira, descrevemos a utilização da abordagem na fonte citada para criação de um código que contempla a extração, tratamento e a arquitetura para armazenamento de dados prontos para análises. Além disso, foram mensuradas algumas métricas que visam avaliar benefícios da abordagem proposta. Durante este processo foi possível aplicar e entender como se manifestam os conceitos de engenharia de software na criação de um data *warehouse*. Resumidamente o trabalho é um exemplo da criação de *data warehouse* com algumas das práticas e requisitos que um projeto de dados pode levar em conta. Conseguimos notar uma melhora em 6% das entradas dos dados referentes a contatos e 75% maior eficiência no armazenamento dos dados.

Palavras-chave: Engenharia de Software, Ciência de Dados, Engenharia de Dados e Bases Públicas.

ABSTRACT

In recent years, the fields of data science and engineering have been consolidating. These fields are important and help extract information and knowledge from the large amount of data available today. The demand for data implies methods for storing data. This work aims to provide a case study of an approach to creating data warehouses that advocate data quality, reusability, efficiency, and reliability. In order to carry out this study, we understood the public data from the National Business Registry as an opportunity for easily accessible data to use the proposed approach. Thus, we describe the use of the approach in the cited source to create a code that includes the extraction, processing, and architecture for storing data ready for analysis. In addition, some metrics were measured to assess the benefits of the proposed approach. During this process, it was possible to apply and understand how software engineering concepts manifest themselves in the creation of a data warehouse. In short, the work is an example of the creation of a data warehouse with some of the practices and requirements that a data project can take into account. We were able to notice a 6% improvement in contact data entries and 75% greater efficiency in data storage.

key-words: Software Engineering, Data Science, Data Engineering and Public Data.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ACRÔNIMOS

BI	Business Intelligence
CNPJ	Cadastro Nacional de Pessoa Jurídica
DE	Data Engineering
DINF	Departamento de Informática
DS	Data Science
DW	Data Warehouse
EPP	Empresa de Pequeno Porte
ES	Engenharia de Software
ETL	Extract, Transform, Load
ELT	Extract, Load, Transform
GDS	Greater Data Science
LAI	Lei de Acesso à Informação
MEI	Micro Empresário Individual
ME	Micro Empresa
RE	Requirement Engineering
SE	Software Engineering
UFPR	Universidade Federal do Paraná

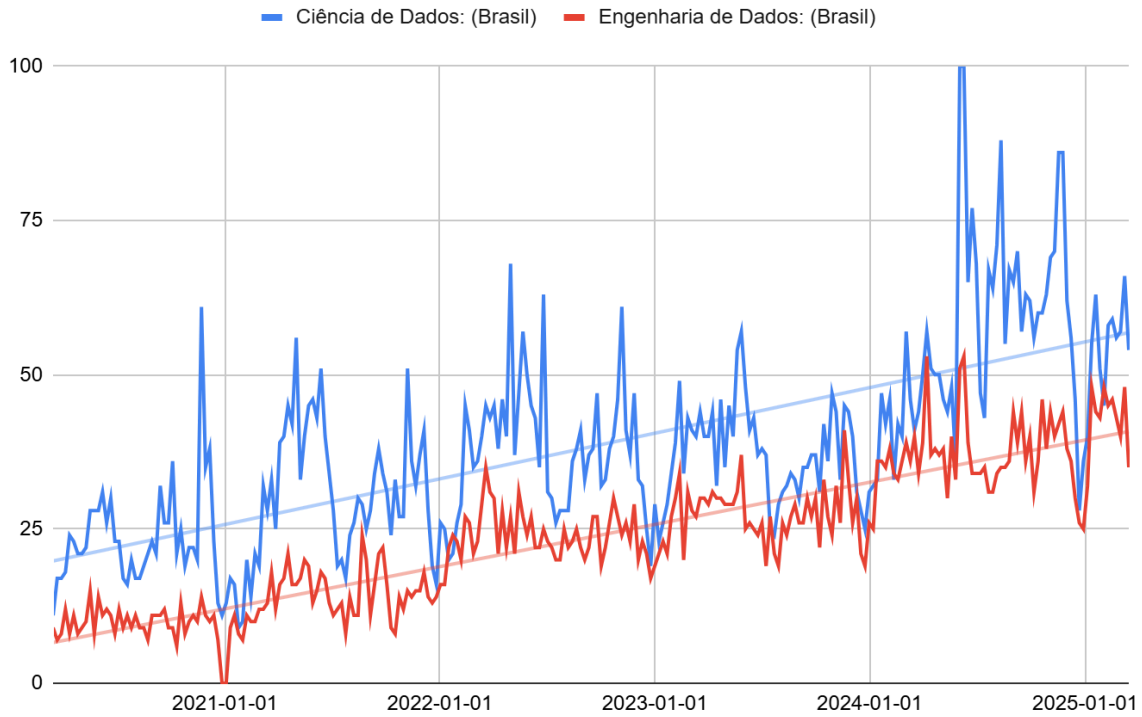
SUMÁRIO

1. INTRODUÇÃO

1.1. Contexto

Nas últimas décadas novos campos de conhecimento vêm se formando. Donoho (2017) faz uma retrospectiva de 50 anos sobre o entendimento do que é Ciência de Dados. Já no livro organizado pelo ENAP (2022) logo em sua apresentação se faz a seguinte afirmação “últimos anos têm vivenciado uma revolução na forma como os dados são produzidos e como circulam entre pessoas, empresas e governos.”. Nesse sentido, entende-se as áreas de ciências de dados e engenharia de dados como áreas importantes. Fazendo uma consulta no Google Trends podemos ver, Figura 1, que nos últimos 5 anos as consultas pelos termos relacionados vêm crescendo.

Figura 1 - Consulta aos termos “Ciência de Dados” e “Engenharia de Dados” nos últimos 5 anos no Brasil.



Fonte: Dados Google Trends. Gráfico autoria própria. Número de buscas pelos termos normalizados.

Ambas as áreas se apresentam por meio de códigos e *softwares*, isto é, devem passar por processos de engenharia de *software*.

1.2. Problema

As áreas de engenharia e ciência de dados são campos de conhecimento importantes que se utilizam de *softwares*. Esses campos na criação de seus projetos têm problemas e necessidades parecidos com outros projetos de *software*. Evidentemente essas áreas também têm suas particularidades. No entanto, acreditamos que para obtenção de bons *softwares* com maior qualidade, que sejam confiáveis, robustos, eficientes e reutilizáveis pode se fazer uso dos conceitos da engenharia de *software*. Atualmente entende-se que 80% do trabalho realizado nos projetos de dados se aplica na parte da consolidação dos dados para análise.

1.3. Objetivo

Utilizando a base de Cadastro Nacional de Pessoas Jurídicas (CNPJs) o objetivo desse trabalho é a utilização de uma abordagem usando conceitos de engenharia de software na criação de *data warehouse* (DW). Seguindo os passos da abordagem fizemos algumas análises e códigos necessários que contemplam desde a extração dos dados brutos da base até a utilização desses dados para gerar a informação. Nesse sentido, este trabalho propõe utilizar engenharia de *software* na parte mais “custosa” da ciência de dados. Isto é, no tratamento e limpeza dos dados. Logo, fornece um exemplo de práticas e preocupações durante o processo de engenharia de dados. Especificamente os principais componentes do desenvolvimento do projeto proposto são:

1. Exploração inicial das bases: Com finalidade de conhecer as bases e entender possíveis utilizações.
2. Planejamento: Evidenciar a utilidade das informações exploradas na parte anterior.
3. Métricas e Qualidade: Verificação e criação de métricas de qualidade para os tratamentos planejados.

Já o código proposto pode ser entendido nos seguintes componentes:

1. *Scripts* que faz *download* do histórico dos arquivos públicos.
2. *Script* que faz a atualização dos arquivos.
3. Tratamentos e validações dos dados.
4. Criação de um *data warehouse*.

Em todos os passos e códigos deverão ser discutidas as boas práticas e conceitos da engenharia de dados.

1.4. Justificativa

Neste trabalho, analisaremos uma base real onde se podem tirar análises e resultados reais. No entanto, o principal objetivo é elucidar as práticas e requisitos na parte mais trabalhosa da área de dados. Mesmo que cada base e projeto tenha sua particularidade entender melhor o processo da criação de *software* dessa área pode revelar requisitos comuns da área no geral. Donoho (2017) destaca:

> "GDS1: Coleta, Preparação e Exploração de Dados" é mais importante do que "GDS5: Modelagem de Dados", conforme medido usando o tempo gasto por praticantes. Porém, houve poucos esforços para formalizar a exploração e limpeza de dados e tais tópicos ainda são negligenciados no ensino. Alunos que apenas analisam dados pré-cozidos não estão tendo a chance de aprender essas habilidades essenciais."(Donoho, 2017).

Nesse sentido, estudos neste tópico "negligenciado" de limpeza, tratamento e exploração dos dados, talvez leve a melhor entendimento e eficiência na criação de códigos dessa parte da Engenharia de dados.

De maneira secundária, o projeto serve como exemplo de um tratamento em uma base pública. Bases públicas são importantes para democracia e denotam e deixam transparecer indicadores importantes dos desempenhos de políticas públicas. Ou seja, este trabalho serve como exemplo do tratamento inicial necessário para extração de informação de fontes públicas.

1.5. Organização do Documento

Na próxima seção são elucidados de maneira um pouco mais aprofundada alguns dos conceitos importantes para o trabalho desenvolvido. Na seção de número três vamos explorar alguns trabalhos correlatos. Utilizando as concepções das partes anteriores, na quarta seção, é desenvolvido a aplicação da abordagem da criação do *data warehouse* em si. Por fim, nos últimos parágrafos do trabalho é discutida a conclusão do trabalho desenvolvido bem como as possibilidades para trabalhos futuros.

2. CONCEITOS BÁSICOS

2.1. Engenharia de *Software*

O objetivo do trabalho é exemplificar e analisar o processo de produção de *software* na área de dados. De maneira mais específica, uma abordagem para a criação da arquitetura de armazenamento. “A disciplina da engenharia que se preocupa com todos os aspectos da produção de *softwares* dos estágios iniciais de especificação até a manutenção do sistema” é a engenharia de *software* (Sommerville, p.7). Ou seja, temos a intenção de trabalhar a engenharia de *software*, porém, devido às limitações do contexto focado mais na parte da especificação do que na manutenção do *software*. Fritz Bauer em Pressman (p.13) define de maneira simplificada a engenharia de *software* com o estabelecimento de práticas que levem a um *software* mais confiável e eficiente.

2.2. Ciência de Dados

Como já citado anteriormente, uma das áreas que vêm se consolidando é a ciência de dados. O desenvolvimento deste trabalho foca mais na parte de engenharia de dados, no entanto, a ciência de dados é uma área correlata. O cientista de dados pode ser visto como um “*stakeholder*” de projetos de engenharia de dados. Entendemos que todo dado é importante a partir da informação que se extrai dele. Igualmente sabemos que bases públicas são bases de interesse da sociedade, entes públicos e privados de maneira geral. Podemos inferir que os usuários dos dados ao tentar extrair informação dos mesmos estão fazendo o papel de cientista de dados. Para melhor entender essa afirmação vamos brevemente descrever alguns conceitos. Qual é a diferença entre “dados”, “informação” e “conhecimento”. Semidão (2014, p.184) faz uma diferenciação entre os termos utilizando várias visões.Trazemos aqui de maneira resumida algumas notas dele sobre os termos no Quadro 1:

Quadro 1 - Dados, Informação e Conhecimento

Termos	Notas
Dados	Elemento primário; isento de significação; número; símbolo; primeira percepção; elemento material; externo à mente; indício; insumo para informação; ligado à tecnologia computacional.
Informação	Reunião de dados; dados processados; agregação de semântica aos dados; conhecimento registrado; insumo para o conhecimento; sinal comunicado; mensagem; nota; notícia; novidade; pré-cognição.
Conhecimento	Informação aplicada em um contexto; informação para tomada de decisão; culminância do processo cognitivo; memória; cabedal de informações na mente; tácito; individual; social; organizacional.

Fonte: Semidão, 2014.

Colocando os termos em “uso” e relacionando-os com a base trabalhada como exemplo. Um conjunto de “dados” é a lista de empresas ativas no Brasil. Uma informação é agregar as empresas ativas por período de tempo e obter o número de empresas ativas no Brasil por períodos de tempo. Já um conhecimento é colocar essa informação em um contexto, como a pandemia ou uma crise econômica e entender se as empresas estão sendo afetadas. Este trabalho tem o objetivo de focar no processo entre dados e informação, parte importante muitas vezes negligenciada. No contexto deste trabalho, o objetivo é transformar dados brutos de CNPJs em dados estruturados prontos para gerar informações úteis para entes públicos ou privados.

O conceito de “Ciência de Dados” se relaciona com a extração de informação. Donoho (2017) faz um estudo de 5 décadas sobre o termo “Ciência de Dados”. Utilizando desse estudo podemos destacar alguns aspectos. Um dos dos entendimentos é que as GDS (Greater Datas Science - Conceito amplos de Ciência de Dados) pode ser divididas em seis atividades:

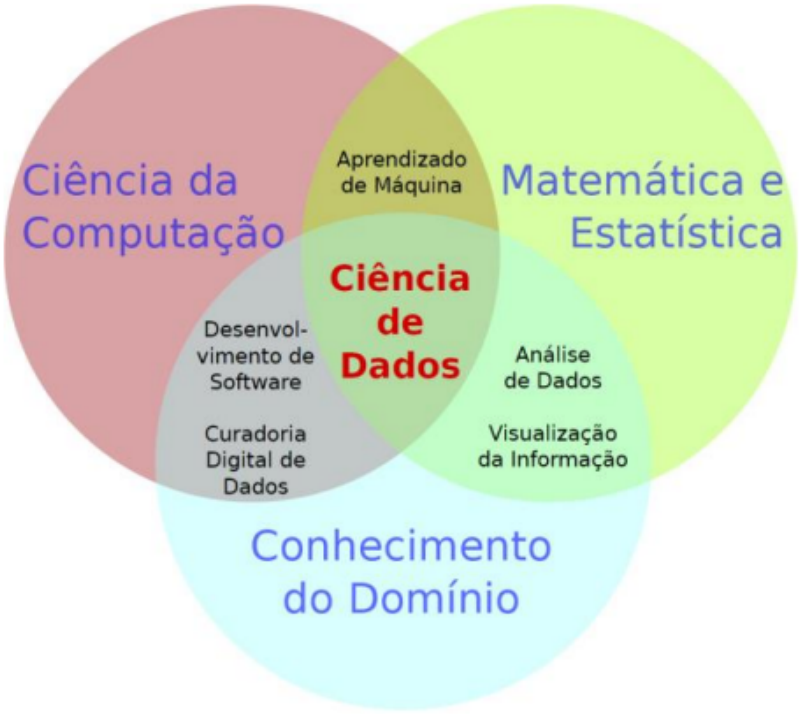
1. Coleta, preparação, e exploração de dados.
2. Representação e transformação de dados.
3. Computação com dados.
4. Modelagem de Dados.
5. Apresentação e visualização de dados.
6. Ciência sobre a ciência dos dados.

As seis atividades são correlatas e podem ser entendidas como “Ciência de Dados”. No entanto, segundo Donoho a categoria que se destaca como representação de ciência de dados é a “Modelagem de Dados”. Essa categoria por sua vez pode ser entendida em duas divisões: “Generativa” e “Preditiva”. Independentemente das divisões e subdivisões do campo, a ciência de dados resumidamente é a ciência de aprender a partir de dados e o estudo de métodos e tecnologias correlatos (Donoho, 2017).

Em um site que o governo brasileiro fomenta o estudo na área, ele utiliza a seguinte definição: “Ciência de dados é uma ciência multidisciplinar que envolve técnicas computacionais, estatísticas e matemáticas, entre outras, com o objetivo de resolver problemas complexos, utilizando para isso grandes conjuntos de dados.” (BRASIL, 2025).

Já Rautenberg (2019) sumariza a ciência de dados no seguinte diagrama de Venn, Figura 2:

Figura 2 - Interdisciplinaridade da Ciência de Dados



Fonte: Rautenberg (2019)

2.3. Engenharia de Dados

Utilizando Reis (p.20) podemos definir engenharia de dados como:

> “[...] é o desenvolvimento, implantação, manutenção de sistemas e processos que partem dos dados brutos e produzem informações consistentes de alta qualidade que suportam os casos de uso posteriores como análises e

> machine learning. Engenharia de dados é a intersecção de segurança, gerenciamento de dados, DataOps, arquitetura de dados, instrumentação e engenharia de *software*. Um engenheiro de dados gerencia o ciclo de vida dos dados, começando com os dados do sistema origem e terminando servindo os dados para os casos de uso como análise ou para aprendizado de máquinas ". (Reis, 2022, p.20).

Os conceitos elucidados aqui não são unívocos. De certa maneira, vemos que eles se sobrepõem em muitos dos casos. No trecho destacado, Reis entende a engenharia de dados como uma **intersecção** de vários campos. Entre estes campos há a "engenharia de *software*". O entendimento não é que a engenharia de dados é uma subconjunto de engenharia de *software*. Nem tão pouco o contrário. Ambos os conceitos são mais amplos em diferentes sentidos. Engenharia de *Software* pode ser aplicada em qualquer *software*, não somente em *softwares* que trabalham com análise de dados. Já a engenharia de dados foca e tem mais particularidades, muitas vezes não tratadas na engenharia de *software*. Além disso, vemos uma sobreposição entre engenharia e ciência de dados. No tópico anterior ressaltamos que Donoho (2017) cita como uma das atividades de ciência de dados "Coleta, preparação, e exploração de dados", sendo que a mesma atividade pode ser entendida como função do engenheiro de dados. A relação entre os três campos é complexa, eles contribuem para o projetos de *software* que trabalham com dados de diferentes maneiras. Neste trabalho entendemos a engenharia de *software* como mais abrangente, permeando assim a engenharia de dados. Já a abordagem proposta e desenvolvida pode ser entendida como um projeto de engenharia de dados. Segundo Reis (2022) um engenheiro de dados coleta os dados, armazena eles e prepara para o consumo de um cientista de dados, analista ou outro.

2.4. Bases Públicas

No ano de 2011 foi sancionada a Lei de Acesso à Informação (LAI) que prevê "divulgação de informações de interesse público, independentemente de solicitações" (Lei nº 12.597/2011). Independentemente de solicitação implica em divulgação espontânea de dados. Essa divulgação possibilita um melhor entendimento da situação do país, ou seja, a transparência de medidas públicas. Essa transparência, segundo Rodrigues (2014), aumenta a eficiência e eficácia de governos, pois faz com que o governo tenha mais cuidado por conta da vigilância das pessoas e entidades. Em outras palavras, é possível, fazer a curadoria e escrutínio de medidas e ações governamentais através de dados públicos.

2.5. Cadastro Nacional de Pessoas Jurídicas

Atualmente, dentre as informações e dados disponibilizadas pelo governo sem solicitação se destacam para esse trabalho os dados do “Cadastro Nacional da Pessoa Jurídica - CNPJ” e os boletins quadrimestrais “Mapa de Empresas”. Entendesse que ambos ajudam a entender as mudanças no panorama empresarial brasileiro. Isto é, são materiais de “interesse público” relevantes e que possuem potencial para embasar ações públicas e/ou privadas.

Os boletins quadrimestrais chamados “mapas das empresas” sumarizam alguns dos dados que podem ser extraídos das bases de cadastro de pessoas jurídicas e complementam com outras informações. Segundo o site onde os boletins estão disponíveis: “O BOLETIM DO MAPA DE EMPRESAS representa a descrição detalhada de dados e informações relevantes sobre o ambiente de negócios e a descrição de ações voltadas a impactar positivamente o cenário econômico” (BRASIL, 2025).

Isto é, além do documento extrair informação das bases de cadastro de cnpjs ele também tem a pretensão de associar as informações com a tomada de medidas do governo. Portanto, essas informações são importantes para sociedade e governo para o entendimento da necessidade da manutenção ou criação de ações para evitar impactos econômicos negativos ou suscitar ampliação de medidas vigentes. No entanto, uma mesma base de dados pode ser tratada de diferentes formas. Isto é, o mesmo dado pode gerar diferentes informações e conhecimentos. Logo, este trabalho tem a intenção de preparar estes dados pensando em *stakeholders* não tão específicos, mas que tenham a intenção de extrair informação que possam ser utilizadas de forma pública ou privada para tomada de decisões. Como ressaltado anteriormente, uma das divisões de Modelagem de dados é a modelagem preditiva. Sendo assim, é importante que a abordagem sustente a ingestão de dados de históricos e mantenha os registros ao atualizar as informações. Além disso, também relevante é deixar disponível outras possibilidades para extração de informações dos dados diferentes das informações presentes dos boletins do governo.

As bases dos CNPJs estão disponíveis em um site do governo. Na página principal do site referente às bases é possível achar a informação que a área técnica responsável é a “RFB - Secretaria Especial da Receita Federal do Brasil”, um e-mail para contato, a periodicidade das bases e a data da última atualização. Além disso, temos 5 links, três deles são pdfs, dos pdfs dois deles são notas técnicas descritivas sobre as bases e o último pdf possui o dicionário de dados das bases. Já os outros dois links levam à páginas onde podem ser feitos os download de arquivos.

A página que mais interessa para as bases do “Cadastro de CNPJs” possui 23 diretórios. A maioria dos diretórios correspondem a um mês de um ano e contém as bases

com as informações disponíveis do próprio mês. No presente estão disponíveis bases de maio/2023 até fevereiro/2025. Por sua vez, os sub-diretórios contém 6 arquivos “zips” complementares: Cnaes.zip, Motivos.zip, Municipio.zip, Naturezas.zip, Paises.zip e Qualificações.zip. E, mais 30 arquivos onde 10 arquivos correspondem às partições da base Empresas, 10 são as partições da base Estabelecimento e, por fim, os arquivos restantes fazem parte da base de sócios. A descrição das tabelas mais detalhadas está na próxima seção 4 com a análise exploratória dos arquivos.

2.6. *Data Warehouse*

Um *Data Warehouse* (armazém de dados) é um sistema que a partir de uma fonte periodicamente extrai dados e os armazena em uma estrutura dimensional normalizada (*Building a Data Warehouse*, p.1). Na prática muitas vezes é visto como uma forma de organizar os dados presentes em um data Lake (REIS, p. 318). Em uso os *data warehouse* mantém informação histórica, normalmente atualizada em lotes, utilizada para análises e inteligência de negócios (BI) (*Building a Data Warehouse*, p.1). A base pública utilizada neste trabalho é atualizada mensalmente e a solução proposta visa armazenar esses dados disponibilizados em lote.

2.7. *Data Lake*

Diferente de um *Data Warehouse* o *Data lake* é um sistema de armazenamento não estruturado. Ao invés, de se fazer o ETL (*Extract, Transform, Load*, em português, Extração, Tratamento e Carga) pode ser feito o ELT (*Extract, Load, Transform*, em português, Extrair, Carregar e Transformar), ou seja, o usuário tem que fazer a transformação na leitura do dado. Não há preocupação na forma de armazenamento da informação. É uma arquitetura que preconiza o acúmulo de informação de várias fontes e em diversos formatos (REIS, p318). Como este trabalho vai trabalhar em somente uma fonte e deseja relatar o processo de tratamento da informação. Logo a abordagem proposta se trata da criação de um *Data Warehouse* e não um *Data Lake*.

2.8. *Práticas de Engenharia de Software utilizadas no projeto*

Para a criação de um data Warehouse iremos contemplar práticas de Engenharia de Software (ES) durante o desenvolvimento. Foram adotadas algumas práticas de qualidade, robustez e organização de projeto.

2.8.1. Versionamento de código

A prática de versionamento de código permite rastreabilidade das alterações, recuperação de versões anteriores e trabalho colaborativo de maneira mais organizada. Foi utilizado o GitHub para este controle de versão e colaboração.

2.8.2. Modularização

Seguindo o princípio da separação de responsabilidade, facilitando a manutenção, leitura e reuso de componentes, foi utilizado o princípio da modularização. Para isso, o código foi dividido dentro das etapas do ETL e mesmo dentro de cada etapa os arquivos estão separados dentro de funções diferentes para cada tratamento de dados.

2.8.3. Validação e levantamento de requisitos

Na etapa de planejamento, foram definidos os requisitos funcionais e não funcionais. Onde os requisitos funcionais descrevem comportamentos esperados do software, como período de atualização. Os requisitos não funcionais abordam aspectos como desempenho, uso eficiente dos recursos e a integridade. Sendo uma etapa da ES, orienta o desenvolvimento para entender as necessidades do *stakeholder*.

2.8.4. Mensuração da Qualidade

A mensuração de qualidade é feita na ES a partir de métricas, as métricas neste projetam serão definidas com base na qualidade dos dados tratados, permitindo uma avaliação dos ganhos dessa abordagem e a importância do tema.

2.8.5. Documentação

Sistemas de Engenharia de Software adotam representações visuais como diagramas e modelos para facilitar o entendimento e compreensão da estrutura, funcionamento e objetivo, adicionando isso a uma documentação escrita formando a documentação do projeto como um todo.

3. TRABALHOS CORRELATOS

Neste capítulo, serão analisados alguns trabalhos correlatos com a área de pesquisa que está sendo abordada neste trabalho. Podemos classificar os trabalhos correlatos em três grupos: trabalhos sobre engenharia de *software*, trabalhos sobre bases públicas e trabalhos em engenharia de dados. Neste sentido, destacar algumas pesquisas que elaboram sobre os temas relacionados.

Sobre engenharia de dados destacam-se dois textos. No trabalho de conclusão de curso, “Construção de um Processo ETL Automatizado em Dados de Campanhas de uma Empresa no Setor Bancário”, Álvaro Martins descreve a criação de um sistema. Este sistema automatizado de tratamento e unificação dos dados permitiu uma melhor visualização das informações e reduziu o tempo de processamento das informações (MACHADO, 2023). Já Ferreira (2010) no artigo “O processo ETL em sistemas *Data Warehouse*” discute sobre o processo de ETL e as ferramentas disponíveis para tratamento de dados e construção de *data warehouses*. Ambos os textos falam sobre ETL e sua importância.

Outro tema relacionado é o uso de bases públicas, sobre este assunto destaca-se o livro: “Ciência de Dados em Políticas Públicas uma Experiência de Formação”. O assunto não é especificamente o uso de dados públicos, mas a ciência de dados em políticas públicas. Logo, independentemente da total ou parcial disponibilidade dos dados utilizados nas pesquisas elencadas no livro, este trata de bases públicas. A obra é um compilado de diversos trabalhos dos alunos do curso de especialização em “Ciência de dados aplicada a Políticas Públicas”. Dessa forma, o livro contempla 11 exemplos do uso da ciência de dados na tomada de decisão. Este trabalho foca principalmente na parte inicial do tratamento de dados de uma base pública. E, não foge do escopo discutir medidas a serem tomadas com os dados tratados. No entanto, contempla o trabalho inconspícuo, mas certamente presente nos trabalhos presentes no livro.

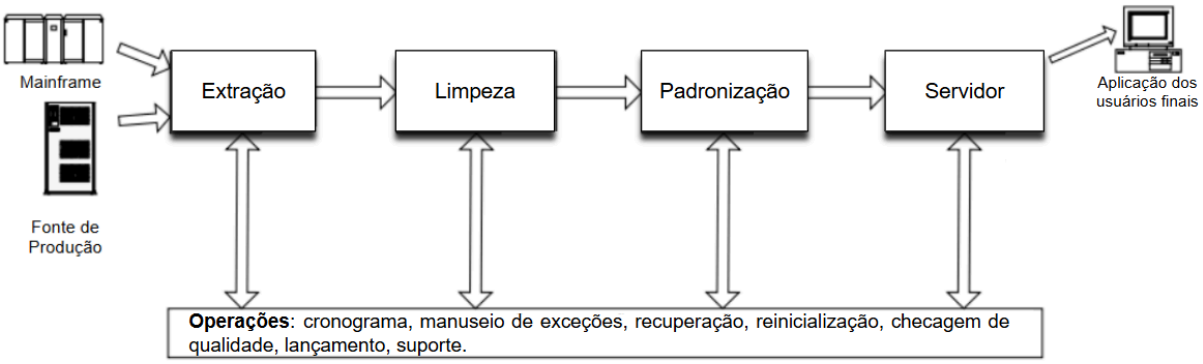
Por último, destaca-se o trabalho “Os efeitos da experiência de desenvolvedor no processo de ensino-aprendizagem de engenharia de *software*”. Neste artigo, Zacarias (2025) discute os aspectos da “Aprendizagem Baseada em Projetos”. Não é um trabalho que trata somente da Engenharia de *Software*, mas de como podemos aprender engenharia de *software* fazendo sua aplicação em projetos. De certa maneira, em um contexto mais específico, de ciência de dados, também é um dos objetivos deste trabalho. Isto é, aplicando os conceitos de ES na construção de um *data warehouse* durante este trabalho, estamos exemplificando e aprendendo ES.

4. ABORDAGEM PROPOSTA

4.1. Modelagem do Sistema

Um *data warehouse* se apresenta de diferentes formas nos mais diversos usos e projetos, no entanto, algumas características se mantêm. Seguem o que podem ser entendidos como os 4 estágios do *data warehouse* representados, na Figura 3, abaixo:

Figura 3 - Estágios *Data Warehouse*



Fonte: KIMBALL (2004, p.18). Traduzido e adaptado.

4.2. Definições Estágios Data Warehouse

Para entender a Figura 3, e posteriormente Figura 04, definiremos os 4 estágios de *Data Warehouse* segundo Kimbal (2004): Extração, Limpeza, Padronização e Armazenamento no Servidor.

O primeiro passo no processo de captura de dados e criação de um *data warehouse* é a extração, segundo Kimball (2004), “Extração é o primeiro passo no processo de obtenção de dados para o ambiente de *data warehouse*. Extrair significa ler e entender os dados da fonte e copiá-los para o sistema ETL para posterior manipulação. Nesse ponto, os dados já pertencem ao *data warehouse*” (Kimball, 2004 p.19)

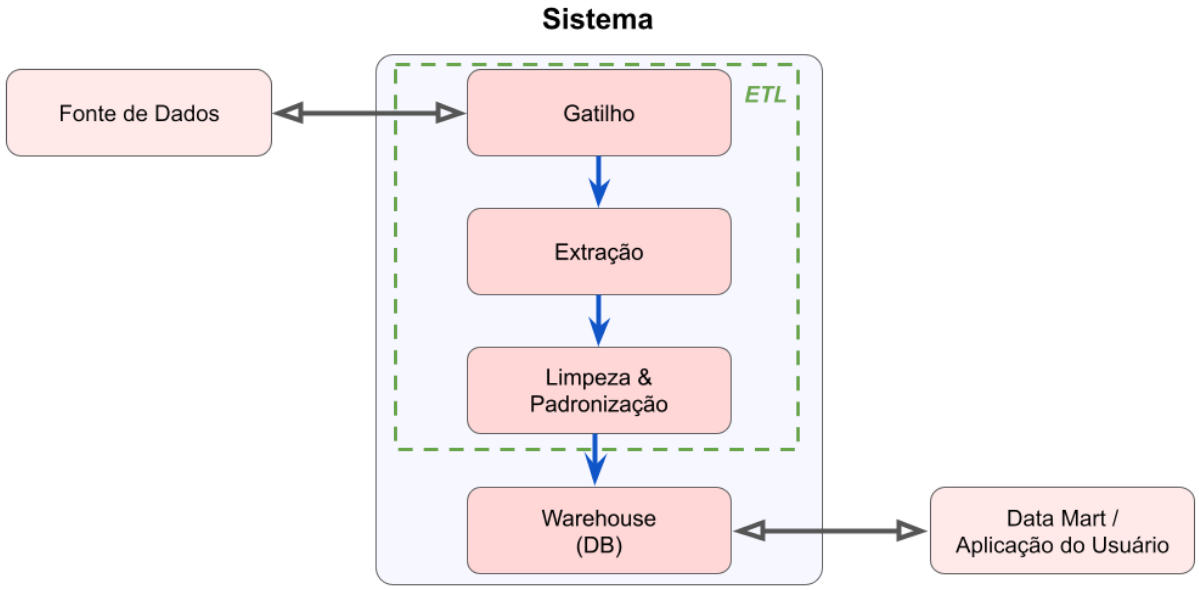
Como próximo passo temos a limpeza dos dados, parafraseando Kimball, são inúmeras as possibilidades de limpeza, como correção de erros de ortografia, resolução de conflitos de informações, lidar com elementos ausentes e analisar o formato adequado dos dados. Isso agrega valor ao *data warehouse*, essas transformações também podem gerar dados para análise de confiabilidade e melhoria na fonte dos dados.

A padronização é considerada a terceira etapa do processo, por vezes, pode se referir a um subpasso da limpeza. Um exemplo é o campo de endereço, onde abreviações como R. se tornam Rua, Ap. é padronizado para Apartamento, SN, S/N, S / N são todas transformadas para sem nome.

Kimball propõem o uso de um servidor ETL, que seria um servidor dedicado para todo este processo, ele seria responsável por realizar as transformações e realizar a carga dos dados aplicando as limpezas e padronizações geralmente definidas pelo negócio. Após isso faria o carregamento dos dados no *data warehouse* para disponibilização.

Como citado na seção 2.6 os *data warehouses* normalmente recebem dados em lote. Logo, a extração deve ocorrer de maneira periódica, então deve existir um mecanismo que inicia à extração. Isto é, deve existir um gatinho para começar as extrações dos dados. Além disso, apesar da padronização e limpeza se referirem a tarefas distintas, entendemos que na prática limpeza e padronização acontecem concomitantemente. Na Figura 4, apresentamos uma adaptação do modelo de Kimball:

Figura 4 - Modelo do Sistema ETL e *Data Warehouse*



Fonte: Autoria própria. Utilizando Google Drawing.

Este delineamento serve como base da arquitetura do código pretendido. No entanto, existem algumas atividades anteriores à própria construção do código relacionadas ao entendimento da base e dos stakeholders ou utilização da informação. Entendemos que a abordagem deve partir do entendimento dos dados disponíveis, relacionar com as possíveis utilizações da base para então a construção da arquitetura em si.

4.3. Requisitos Genéricos de *Data Warehouses*

4.3.1. Requisitos Funcionais

Requisitos funcionais são os requisitos que descrevem o que o sistema deve fazer (Somerville, p.85). No Quadro 2 são elencados quatro requisitos funcionais de *data warehouses* de maneira genérica.

Quadro 2 - Requisitos Funcionais Genéricos

#	Requisito Funcionais	Motivação
01	O *Data Warehouse* deve estar atualizado.	A cadência de atualização de dados vai depender da utilização. No entanto, os dados têm utilidade mediante à um período específico.
02	O *Data Warehouse* deve permitir consultar dados históricos.	Para modelagem preditiva e para comparações entre diferentes períodos é necessário a possibilidade de retroagir as informações.
03	O *Data Warehouse* não deve permitir duplicidade de dados.	Um dado duplicado só consome espaço do hardware e não traz benefício algum para o usuário.
04	Os dados obtidos através do *Data Warehouse* devem ser válidos.	Os dados devem ser verídicos e coerentes. Ou seja, as informações devem ser o mais próximo da realidade e de maneira não conflitante.

Fonte: Autoria própria.

4.3.2. Requisitos Não funcionais

Requisitos não funcionais dizem respeito a características desejadas do sistema. Não diretamente preocupada com os serviços prestados aos usuários (Somerville, p.87). No Quadro 3 são elencados três requisitos não funcionais de *data warehouses*.

Quadro 3 - Requisitos Não Funcionais

#	Requisito Não Funcionais	Motivação
01	O *Data Warehouse* deve otimizar o espaço utilizado.	Tanto para soluções "On-premise" quanto para soluções em Cloud. O uso desnecessário de espaço representa custo desnecessário.
02	O *Data Warehouse* deve ser rápido.	Apesar da velocidade dependem do ambiente onde o dado é disponibilizado. A arquitetura dos dados podem onerar o sistema.
03	O *Data Warehouse* deve possuir dados de qualidade.	Na medida do possível os dados devem ser fidedignos.

Fonte: Autoria própria.

4.4. Resumo da Abordagem

Podemos entender os passos para abordagem como os passos para montar a arquitetura da seção 4.1 levando em consideração os requisitos de 4.2. Dessa maneira, definimos à abordagem em 5 passos:

1. **Entendimento dos dados.** Consiste em descobrir ou interpretar os dados disponibilizados. Entender o significado da informação em um registro, antes de entender a possível sumarização do mesmo. Exemplo seção 5.2.
2. **Análise Exploratória.** Classificação dos dados na sua natureza, checar completude/preenchimento, checar domínio e variação das variáveis. Exemplo seção 5.3.

3. **Relacionamento dos dados.** Entender a correlação entre as fontes e/ou campos disponíveis. Podendo servir para o diagrama de entidade e relacionamento. Exemplo seção 5.4.
4. **Entender utilidade e *stakeholders*.** O consumo da informação deve ser feito com algum objetivo. Dados sem utilidades não devem ser levados ao *data warehouse*. É necessário entender os usuários e o uso deles para os dados. Exemplo seção 5.5.
5. **Construção efetiva do *data warehouse*.** Isto é, construção dos elementos descritos na Figura 04. Exemplo seção 5.6.

No próximo capítulo aplicamos a abordagem na fonte de dados exemplo do estudo de caso.

4.5. Definição de Métricas

Para validação da abordagem foram estipuladas algumas métricas. Com a utilização dessas métricas conseguimos indicar o benefício da utilização da abordagem. As métricas utilizadas foram:

- **Qualidade dos Dados:** Limpeza de dados inválidos ou que não fazem sentido. Acreditamos que para o usuário final da base não é relevante dados incompletos ou inutilizados. Em outras palavras, é melhor ficar com Nulos do que um campo da tabela invalido e sem significado.
- **Consistência:** Os dados não deveriam se contradizer. Alguns campos se relacionam entre si, mas vemos que o dado é inconsistente levando a múltiplas interpretações. Entendemos que existe uma hierarquia nos dados que podemos corrigir as inconsistências com o campo mais confiável. Ou ainda, caso não haja campo mais confiável descartando a informação. Novamente partindo da premissa que é melhor não ter informação do que ter uma errada.
- **Armazenamento:** Apesar de atualmente ser amplamente utilizado em sistemas em nuvem com capacidade de armazenamento cada vez maiores, todo uso de capacidade tem um custo. Dessa maneira, sem perder informação analisaremos a redução de uso de armazenamento utilizando formatos adequados.

4.6. Materiais e Métodos

Para a aplicação da abordagem desenvolvimento do código utilizamos algumas tecnologias. Para criação dos diagramas utilizamos o *Lucidchart*, para o desenvolvimento do código em conjunto utilizamos o GitLab do departamento de informática (Dinf) da UFPR. Por fim, para criação do *data warehouse* utilizamos apache spark com pyspark. Sendo que apache e spark é um framework muito utilizado para processamento de dados. Pyspark é a biblioteca que permite códigos Python interagir com o Spark.

5. ESTUDO DE CASO

5.1. Procedimentos da Análise

O objetivo deste estudo de caso é exemplificar a abordagem proposta (seção 04), que utiliza os conceitos de ES, para criação de um *data warehouse*. Logo, nesta seção será explanada a criação do *data warehouse* levando em conta tais conceitos. Desta maneira começamos com uma análise exploratória das bases públicas de cadastro de CNPJ a fim de se familiarizar com os dados disponíveis. Em seguida, pensando no planejamento do *data warehouse*, criamos algumas personas e os requisitos que o sistema deve cumprir. Por último, mensuramos algumas métricas que se relacionam com a qualidade de um *data warehouse*. A base analisada tem uma boa qualidade. Isto é, não se esperava encontrar muitos erros. O objetivo é aplicar conceitos de qualidade da abordagem independente do tamanho do impacto. E, em um suposto uso da arquitetura sugerida, mesmo melhorias pequenas podem representar algum ganho operacional.

5.2. Entendimento dos dados

Na seção 2.5 descrevemos o estado atual do portal onde achamos os arquivos do Cadastro Nacional de Pessoa Jurídica. Nesse tópico vamos aprofundar o entendimento das bases, suas relações e fazer análises exploratórias das bases. Isto é, os três primeiros passos da abordagem: **Entendimento dos dados, Relacionamento dos dados e Análise Exploratória,** estão elencados nos tópicos 5.2 e 5.3. Os arquivos contidos nas partições não possuem cabeçalho e as entradas são separadas por ponto e vírgula. Já as informações estão todas entre aspas, isto é, como se o tipo de todos os dados fossem textuais (*strings*). Os arquivos podem ser separados em dois tipos: arquivos que relacionam código com nome/descrição e arquivos mais complexos com mais campos.

São 6 tabelas simples: Países, Municípios, Natureza Jurídica, Motivo Situação, Qualificação dos Sócios e Cnaes. O layout dessas tabelas estão representados nos Quadros 4 à 9 são os seguintes:

Quadro 4 - Países

Campo	Descrição
Código	Código do País
Descrição	Nome do País

Fonte: Portal de Dados Abertos. cnpj-metadados.

Quadro 5 - Municípios

Campo	Descrição
Código	Código do Município
Descrição	Nome do Município

Fonte: Portal de Dados Abertos. cnpj-metadados.

Quadro 6 - Naturezas Jurídicas

Campo	Descrição
Código	Código da natureza jurídica
Descrição	Nome da natureza jurídica

Fonte: Portal de Dados Abertos. cnpj-metadados.

Quadro 7 - Motivo Situação

Campo	Descrição
Código	Código do Motivo
Descrição	Descrição do Motivo

Fonte: Portal de Dados Abertos. cnpj-metadados.

O layout do Motivo da Situação não consta no arquivo de layout, mas pode ser presumido a partir da base.

Quadro 8 - Qualificação de Sócios

Campo	Descrição
Código	Código da qualificação do sócio
Descrição	Nome da qualificação do sócio

Fonte: Portal de Dados Abertos. cnpj-metadados.

Quadro 9 - CNAES

Campo	Descrição

Código	Código da atividade econômica
Descrição	Nome da atividade econômica.

Fonte: Portal de Dados Abertos. cnpj-metadados.

Todas as tabelas simples, Quadro 2 ao 7, tem o layout simples com duas informações: “Código” e “Descrição”. São dicionários para os códigos presentes em outros arquivos.

Já as tabelas mais complexas são 4: Empresas, Estabelecimentos, Dados do Simples e Sócios.

No Quadro 10 é possível observar o layout da “tabela” Empresas com 6 campos:

Quadro 10 - Layout base Empresas

Campo	Descrição
CNPJ Básico	Número base de inscrição no CNPJ (Oito primeiros dígitos do CNPJ).
Razão Social / Nome Empresarial	Nome Empresarial da Pessoa Jurídica
Natureza Jurídica	Código da Natureza Jurídica
Qualificação do Responsável	Qualificação da pessoa física responsável pela empresa
Capital Social da Empresa	Capital Social da Empresa
Porte da Empresa	Código do Porte da Empresa: 00 – Não Informado 01 - Micro Empresa 03 - Empresa de Pequeno Porte 05 - Demais
Ente Federativo Responsável	O ente federativo responsável é preenchido para os casos de órgãos e entidades do grupo de natureza jurídica 1xxx para as demais naturezas , este atributo fica em branco.

Fonte: Portal de Dados Abertos. cnpj-metadados.

Já no Quadro 11 é possível observar o layout da “tabela” Estabelecimento com 30 campos:

Quadro 11 - Layout base Estabelecimento

Campo	Descrição

CNPJ Básico	Número base de inscrição no CNPJ (Oito primeiros dígitos do CNPJ).
CNPJ Ordem	Número do estabelecimento de inscrição no CNPJ (do nono até o décimo segundo dígito do CNPJ).
CNPJ DV	Dígito verificador do número de inscrição no CNPJ (dois últimos dígitos do CNPJ).
Identificador Matriz/Filial	Código do Identificador matriz/filial: 1 – Matriz 2 – Filial
Nome Fantasia	Corresponde ao nome fantasia
Situação Cadastral	Código da Situação Cadastral: 01 – Nula 02 – Ativa 03 – Suspensa 04 – Inapta 08 – Baixada
Data Situação Cadastral	Data do Evento da Situação Cadastral
Motivo Situação Cadastral	Código do Motivo da Situação Cadastral
Nome da Cidade no Exterior	Nome da cidade no exterior.
País	Código do País
Data de Início Atividade	Data de Início Atividade
CNAE Fiscal Principal	Código da Atividade econômica principal do estabelecimento.
CNAEs Fiscais Secundário	Código da(s) Atividade(s) econômica(s) secundária(s) do estabelecimento.
Tipo de Logradouro	Descrição do tipo de logradouro
Logradouro	Nome do logradouro onde se localiza o estabelecimento.
Número	Número onde se localiza o estabelecimento. Quando não houver o preenchimento do número haverá S/N.
Complemento	Complemento para o endereço de localização do estabelecimento
Bairro	Bairro onde se localiza o estabelecimento.
CEP	Código de endereçamento Postal Referente ao logradouro no qual o estabelecimento está localizado

UF	Sigla da unidade da federação em que se encontra o estabelecimento
Código Município	Código do município de jurisdição onde se encontra o estabelecimento
DDD 1	Contém o DDD 1
Telefone 1	Contém o número do telefone 1
DDD 2	Contém o DDD 2
Telefone 2	Contém o número do telefone 2
DDD do Fax	Contém o DDD do Fax
Fax	Contém o número do Fax
Correio Eletrônico	Contém o e-mail do contribuinte
Situação Especial	Situação especial da empresa
Data da Situação Especial	Data em que a empresa entrou em situação especial

Fonte: Portal de Dados Abertos. cnpj-metadados.

Em seguida, no Quadro 12 é possível observar o layout da “tabela” Dados do Simples com 7 campos:

Quadro 12 - Dados do Simples

Campo	Descrição
CNPJ Básico	Número base de inscrição no CNPJ (Oito primeiros dígitos do CNPJ).
Opção pelo simples	Indicador da existência da opção pelo simples. S - Sim N - Não Em branco - outros
Data de opção pelo simples	Data de opção pelo simples
Data de exclusão do simples	Data de exclusão do simples
Opção pelo MEI	Indicador da existência da opção pelo MEI. S - Sim N - Não Em branco - outros
Data de opção pelo	Data de opção pelo MEI

MEI	
Data de exclusão do MEI	Data de exclusão do MEI

Fonte: Portal de Dados Abertos. cnpj-metadados.

Por último, no Quadro 13 é possível observar o layout da “tabela” Sócios com 11 campos:

Quadro 13 - Sócio

Campo	Descrição
CNPJ Básico	Número base de inscrição no CNPJ (Oito primeiros dígitos do CNPJ).
Identificador do Sócio	Código do Identificador de sócio: 1 – Pessoa Jurídica 2 – Pessoa Física 3 – Estrangeiro
Nome do Sócio (no caso de PF) ou Razão Social (no caso PJ)	Nome do sócio pessoa física ou razão social e/ou nome empresarial da pessoa jurídica e/ou nome do sócio/razão social do sócio estrangeiro
CNPJ/CPF do sócio	CPF ou CNPJ do sócio (sócio estrangeiro não tem essa informação).
Qualificação do Sócio	Código da qualificação do sócio.
Data de entrada sociedade	Data de entrada na sociedade
País	Código do país do sócio estrangeiro
Representante Legal	Número do CPF do representante legal
Nome do representante	Nome do representante legal.
Qualificação do representante legal	Código da qualificação faixa Etado representante legal.
Faixa etária	Código correspondente à faixa etária do sócio

Fonte: Portal de Dados Abertos. cnpj-metadados.

A quantidade de dados é de aproximadamente 5.9 Gb por mês, são 22 meses de histórico somando 129 Gb de dados compactados. Os meses de Agosto/2023 e Dezembro/23 tem alguns arquivos a mais e com nomes diferentes dos convencionais. As outras 20 safras têm o mesmo padrão.

Ainda no contexto da análise exploratória, mas descobrindo aspectos importantes para as próximas etapas dado os layouts verificamos as volumetrias dos dados das tabelas brutas. Começando com as tabelas menores segue a volumetria das tabelas na Tabela 1:

Tabela 1 - Volumetria Tabelas Simples mês 03/2025

Tabela	Volumetria
Municípios	5.572
Países	255
Cnaes	1.359
Motivos	61
Natureza	90
Qualificações	68

Fonte: Autoria própria. Usando Lucid Chart.

Essas tabelas são simples de apenas duas colunas. Verificamos que não existem duplicadas e que sofrem pouquíssima alteração nos meses disponíveis. A volumetria no quadro acima é referente ao mês de fevereiro de 2025. No entanto, nenhuma dessas bases variou de volumetria desde maio de 2023. Com exceção a tabela de Municípios que obteve um novo registro no último mês. Exemplo com as primeiras linhas de cada tabela no Anexo I deste documento. As demais tabelas são mais complexas e portanto serão exploradas nas subseções seguintes.

5.3. Análise Exploratória Base Estabelecimento Completude

As tabelas com mais de dois campos podem ter suas colunas divididas nas seguintes categorias: chaves, identificadores, campos numéricos e campos categóricos. As chaves das tabelas não devem ter nulos. Os campos de identificadores podemos analisar o preenchimento, os campos categóricos podemos analisar a frequência das categorias e os campos numéricos algumas estatísticas.

As tabelas de estabelecimentos são os arquivos com mais informações. Não existe nulos nos três componentes da chave primária: “cnpj_corpo”, “cnpj_filial” e “cnpj_controle” também praticamente não temos nulos nas colunas “id_matriz_filial”, “cd_sit_cad”, “cd_motivo” e “adrs_uf ”. No entanto, destacamos a completude das demais colunas na Tabela 2 a seguir:

Tabela 2 - Percentual preenchimento colunas Estabelecimentos

Coluna	Preenchido (%)
Nome Fantasia	38,65%
Situação Cadastral	100,00%
Motivo Da Sit. Cad.	100,00%
Nome Cidade Exterior	0,06%
Código País	2,21%
Data Início de Ativ.	100,00%
Cnae Primário	100,00%*
CNAE Secundário	45,98%
Tipo Logradouro	98,59%
Logradouro	100,00%
adrs número	100,00%*
adrs complemento	41,74%
adrs bairro	99,42%
adrs cep	99,75%
adrs uf	100,00%
cód. Município	100,00%
cont_ddd1	80,96%
cont telefone 1	80,96%
cont ddd2	8,27%
cont telefone 2	8,26%
cont ddd fax	11,93%
cont fax	11,94%
e-mail	66,11%
Cód. Sit. Esp.	0,05%
Data Sit. Esp	0,05%

Fonte: Autoria própria.

* O cnae Primário possui 2,77% de preenchimento com o CNAE 8888888, que é inexistente oficialmente, servindo como código genérico.

* O Número contém 9,85% de registros constatados como S/N (Sem Número). No Brasil realmente temos endereços onde não se contém números, não conseguimos bases de comparação para saber a corretude desse valor.

Podemos ver uma grande diferença de preenchimento entre os dados de endereço, contato, dados básicos, como CNAE, e presença no exterior de cada CNPJ. Essa diferença de preenchimento e conceito é um indicativo de que é mais vantajoso separar essa tabela em outras tabelas menores. O campo nome cidade exterior, por exemplo, é nulo em praticamente toda tabela. É melhor deixar as tabelas exteriores em outra tabela. Já os

campos de “fax” além de serem pouco preenchidos tem pouco uso atualmente. Logo, são registros que podemos deixar de utilizar. O DDD do telefone sem o número do telefone é inútil e deveriam ser campos retirados também.

5.3.1. Análise Exploratória Base Estabelecimento Frequência Categóricas

Nas variáveis, categóricas, analisamos também a frequência dos valores. Ao analisar a variabilidade de uma variável podemos inferir sobre a relevância da informação ou até mesmo a arquitetura do *warehouse*. Por exemplo, se determinada coluna tivesse 99,99% de preenchimento e somente 0,01% contivesse informação diferente, talvez seja mais interessante deixar essa “lista” menor separada em uma tabela à parte. Logo, nos parágrafos seguintes vamos apresentar um racional seguido da tabela das frequências observadas para algumas colunas do arquivo Estabelecimentos.

Na tabela 3 é possível observar a distribuição entre Matriz e filial e que a grande maioria das não abrem filiais.

Tabela 3 - Distribuição Categorias Matriz/Filial

Matriz/Filial	%
1 - Matriz	95,27%
2 - Filial	4,73%

Fonte: Autoria própria.

Para melhor entendimento da próxima tabela seguem as interpretações de cada uma das situações cadastrais possíveis:

- Situação Nula: acontece por uma inconformidade de dados, duplicidade de inscrição ou suspeita de fraude e não pode ser revertido a empresa ativa.
- Situação Ativa: a empresa está em conformidade com os débitos e impostos e não existem problemas judiciais.
- Situação Suspensa: existe algum ponto das obrigações não foram cumpridos, o envio de declarações, inconsistências nos dados da Receita Federal ou a empresa está sendo investigada por suspeita de fraude, a situação permite regularização e atualização para a situação Ativa.
- Situação Inapta: acontece quando a empresa não cumpre suas obrigações por dois anos consecutivos, muito parecido com a suspensa porém agora com um prazo de 2 anos de suspensão, ainda sendo possível reativar o CNPJ para Ativo.

- Situação Baixada: ocorre quando o próprio empreendedor solicita o fechamento da empresa ou quando por 5 anos ela não apresenta as informações e declarações necessárias, não sendo possível reativar a empresa.

Com a distribuição de situações cadastrais, tabela 04, vemos uma grande concentração em empresas baixadas e ativas. Relevante notar que 46,26% da base não pode ser alterada, por estar baixada ou nula. Portanto, as atualizações poderiam ser feitas somente com as empresas com situação ativa, suspensa ou inapta. E, uma pequena atualização mensal de novas empresas baixadas e nulas para apendar na base de empresas antigas.

Tabela 4 - Distribuição Categorias Situação Cadastral

Situação Cadastral	%
1 - Nula	0,15%
2 - Ativa	39,04%
3 - Suspensa	0,41%
4 - Inapta	14,29%
8 - Baixada	46,11%

Fonte: Autoria própria.

Já a tabela 05 mostra a frequência dos 6 motivos mais frequentes da situação cadastral tem de um total de 61 motivos diferentes. O motivo 0 aparece para todas as empresas onde a situação é Ativa, e todas as outras situações cadastrais tem um motivo diferente do 0 (Sem Motivo), os 2 maiores ofensores sendo Extinção voluntária onde o sócio pede a baixa do CNPJ e Omissão de declaração, onde a baixa ocorre pela falta de declarações por parte dos responsáveis. Os 3 primeiros motivos já somam 88,39 % da base inteira.

Tabela 5 - Distribuição Categorias Código Motivo (6 mais frequentes)

CD MOTIVO	%
0 - Sem Motivo	39,04%
1 - Extinção Voluntária	35,11%
63 - Omissão de Declaração	14,24%
71 - Inaptidão	5,44%
67 - Registro Cancelado	2,12%
73 - Omissão Contumaz	1,50%

Fonte: Autoria própria.

A Categoria do Cnae principal reflete o ramo de atividade principal da empresa, tendo hoje 1360 CNAEs diferentes. Na tabela 06 trouxemos novamente as 6 categorias mais frequentes. A coluna em si possui preenchimento para 100% da base, o destaque maior fica para o código “8888888” que se refere a “Atividade Econômica não informada” o que poderia significar uma falta de informação na base sobre a atividade principal daquela empresa, representando 2,77% da base o que gera 1,8 Milhão de CNPJs sem esta informação.

Tabela 6 - Distribuição Categorias CNAE Principal (6 mais frequentes)

CNAE Principal	%
4781400	5,36%
9492800	5,07%
5611203	2,92%
8888888	2,77%
9602501	2,75%
4712100	2,53%

Fonte: Autoria própria.

Como última variável analisada do arquivo estabelecimentos temos a distribuição através dos estados, tabela 07. Os principais estados sendo São Paulo, Minas Gerais, Rio de Janeiro e Rio Grande de Sul, a concentração por região fica com 50,34% na região Sudeste, 18,46% na região Sul, 17,28% no Nordeste, 8,52% no Centro Oeste e 5,12% na região Norte, 0,25% ainda se concentram em “EX” que se refere Exterior.

Tabela 7 - Distribuição Estados

UF	%
AC	0,23%
AL	0,96%
AM	1,06%
AP	0,22%
BA	5,01%
CE	2,83%
DF	1,72%
ES	2,10%
EX	0,25%
GO	3,56%

UF	%
MA	1,51%
MG	10,90%
MS	1,34%
MT	1,90%
PA	2,03%
PB	1,28%
PE	3,00%
PI	0,87%
PR	6,78%
RJ	8,50%
RN	1,15%
RO	0,70%
RR	0,20%
RS	6,92%
SC	4,76%
SE	0,67%
SP	28,84%
TO	0,68%

Fonte: Autoria própria.

Olhando apenas essas frequências podemos ter uma ideia se os dados estão coerentes e até tirar algumas conclusões. É intuitivo São Paulo ser um dos estados com maior concentração de empresas. Também parece intuitivo que o CNAE mais representativo é o CNAE de um setor de varejo que engloba alguns tipos de comércio. Além disso, conseguimos notar que apenas 41% da tabela se trata de empresas ativas (situação cadastral 2).

5.3.2. Análise Exploratória Base Empresas

De maneira similar a tabela de estabelecimento tiramos algumas métricas do arquivo empresas.

Na tabela 8, analisamos a coluna Natureza Jurídica da empresa que se refere a como o estado classifica legalmente a empresa, temos ao todo 90 naturezas Jurídicas e trouxemos as 5 mais frequentes, que representam 95,91% da base, em ordem elas significam: Empresário Individual, Sociedade Empresária Limitada, Candidato a Cargo Político Eletivo, Associação Privada e Sociedade Simples Limitada.

Tabela 8 - Distribuição Categorias Natureza Jurídica (5 primeiros)

Cd Natureza Jurídica	%
2135	63,88%
2062	23,74%
4090	4,73%
3999	2,25%
2240	1,31%

Fonte: Autoria própria.

A Qualificação do Responsável, tabela 9, se refere a como o responsável por aquela empresa deve ser tratado, tendo 68 qualificações diferentes, as mais comuns são respectivamente: Empresário, Sócio-Administrador, Candidato a Cargo Político Eletivo, Presidente e Produtor Rural. Estes representam ao todo 96,83% da base.

Tabela 9 - Distribuição Categorias Qualificação Responsáveis (5 primeiros)

Cd. Qualificação Responsáveis	%
50	63,86%
49	24,33%
51	4,73%
16	2,92%
59	0,99%

Fonte: Autoria própria.

Já porte da empresa, tabela 10, refere se ao faturamento anual declarado pela empresa no cadastro ou atualização do cadastro, onde Microempresa é até 360 Mil de faturamento anual, Empresa de pequeno porte é de 360 a 4,8 Milhões de Faturamento e Demais se refere a empresas de médio ou grande porte onde o faturamento declarado é maior. Ressaltando que o faturamento aqui é o declarado e não calculado diretamente pela Receita Federal, o que torna a informação menos confiável.

Tabela 10 - Distribuição Categorias Porte Empresas

Porte	%
01 - Microempresa	73,65%
05 - Demais	23,34%
03 - Empresa Pequeno Porte	2,90%
0 - Não informado	0,11%

Fonte: Autoria própria.

O Capital Social da empresa também é preenchido diretamente pelo responsável na hora do cadastro ou atualização, com um valor não pré-definido e contínuo. Logo, por isso, trouxemos um resumo geral dos valores na tabela 11. Sobre os valores, o máximo é R$ 999.999.999.999,00, os valores estão concentrados abaixo dos 10 mil reais para 75% da base. Porém, do outro lado, existem valores muito altos. Os valores de 1 trilhão contemplam 124 empresas, sendo a maioria delas, Holdings, incorporações e participações, onde a natureza dessa empresa é que o capital social seja atrelado ao valor do bem ao qual aquela holding está sendo referenciada. Provavelmente este é o motivo desses valores elevados. No entanto, por ser preenchido diretamente pelo responsável sem uma verificação de corretude por parte da Receita Federal o dado traz menos confiança sobre sua veracidade.

Tabela 11 - Resumo Distribuição Capital Social

Sumário Capital Social	
Preenchido	62.085.952
média	4,51 mi
Desvio Padrão	1,63 bi
mínimo	0 reais
Q1	0 reais
Q2	1 mil
Q3	10 mil
Máximo	1 trilhão

Fonte: Autoria própria.

Apesar de ser um valor contínuo, achou-se oportuno verificar os 5 valores mais recorrentes na base, tabela 12, os quais juntos somam 62,91% da base. Interessante notar que são 28,19% de valores iguais a 0 e 6,22% de valores iguais a 1 Real. Dessa maneira, é possível notar a baixa confiabilidade do dado, onde uma empresa com capital de 1 real não

conseguiria operar verdadeiramente sem condições de pagar taxas, emissão de nota, contratar contador, adquirir bens, pagar fornecedores tendo apenas 1 real de capital.

Tabela 12 - Distribuição Capital Social (5 primeiros)

Capital Social	%
R$ 0,00	28,19%
R$ 1.000,00	11,32%
R$ 5.000,00	9,43%
R$ 10.000,00	7,75%
R$ 1,00	6,22%

Fonte: Autoria própria.

5.3.3. Análise Exploratória Base Simples/MEI

De maneira similar ao arquivo de estabelecimento tiramos algumas métricas do arquivo simples/MEI. O simples nacional se trata de um regime tributário criado em 2006, ele unifica o pagamento de diversos tributos federais, estaduais e municipais, com isso as alíquotas são reduzidas e simplificadas, é um programa voltado a empresas com faturamento até 4,8 Milhões de reais anuais. Notamos, na Tabela 13, que mais da metade da base está contemplada pelo simples. Porém, este arquivo possui 42,7 Milhões de registros (cnpjs), diferente dos 65,2 Milhões de registros de um lote do arquivo estabelecimentos. Logo, esses números podem trazem a análise de que existem 25,4 Milhões de estabelecimentos ativos e 22,9 Milhões de registros como simples. Está análise seria mais interessante levando em consideração quais dos quase 23 Milhões estão com situação de estabelecimentos ativos, mas dá uma noção da relevância do simples.

Tabela 13 - Distribuição Categorias Simples

Opção Simples	%
Sim	53,65%
Não	46,35%

Fonte: Autoria própria.

Na Tabela 14, onde demonstramos a distribuição dos MEIs, notamos que a menor parte da base é considerada MEI, somente 15,8 Milhões de estabelecimentos. O Mei foi criado em 2009, com foco na formalização dos trabalhadores autônomos com faturamento de até 81

Mil reais. MEI é a modalidade de empresa mais simples e garante cobertura previdenciária e pagamento de impostos fixos mensais, diminuindo a burocracia e facilitando o acesso à aposentadoria.

Tabela 14 - Distribuição Categorias MEI

Opção MEI	%
Não	62,95%
Sim	37,05%

Fonte: Autoria própria.

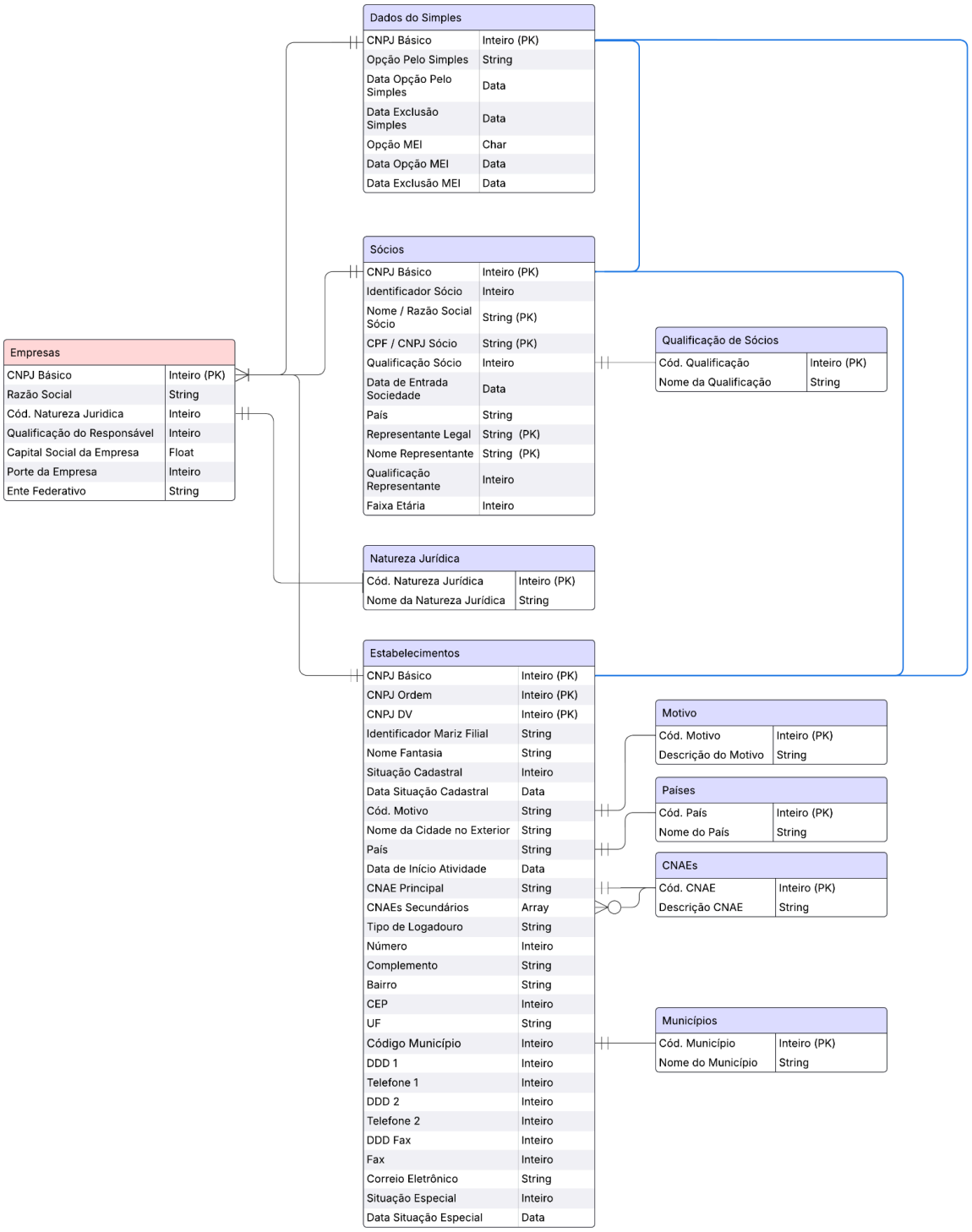
A análise das categorias é importante para validar a importância dessas tabelas. Porém, estes dados soltos não trazem nenhum conhecimento por si só. Caso existissem colunas com muito pouco ou sem preenchimento poderíamos sugerir retirar a coluna ou outro tratamento.

5.4. Relacionamento dos dados

As tabelas se relacionam e o banco de dados “implicitamente” formado por elas está na terceira forma normal. Isto é, não tem nenhum atributo sendo definido em uma tabela por outro atributo que não a chave primária. Talvez pudessem argumentar que CEPs determinam ruas, Cidades determinam estados e/ou CPFs determinam nomes. No entanto, existem cidades homônimas e atualmente os CEPs podem encapsular mais do que apenas uma rua. Por fim, os CPFs estão anonimizados, isto é, não conseguimos saber todos os caracteres do CPF do sócio. Logo, o nome do sócio, ao invés de ser determinado pelo CPF, ele em conjunto com o CPF ajuda a compor a chave da tabela. Sobre a tabela de sócios ainda é importante se notar que uma tupla {cpf, nome} pode ser sócio de mais de uma empresa. No entanto, podemos entender como chave primária da base a composição CNPJ empresa, cpf e nome, ou seja, a tripla {cnpj, cpf, nome}. Essa relação só pode aparecer uma vez na tabela e diz respeito a apenas uma empresa.

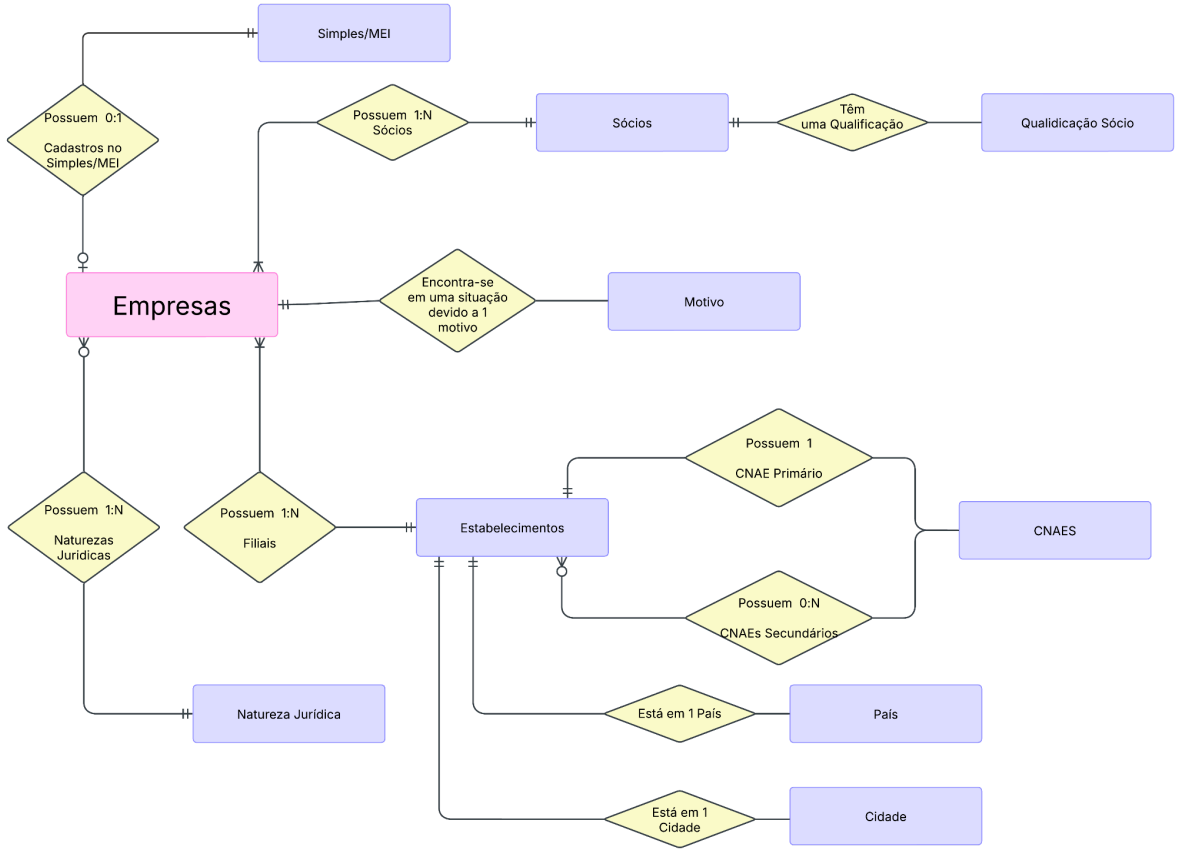
Os arquivos disponibilizados no site do governo são apenas arquivos texto. No entanto, os campos têm nomes e formatos que sugerem uma relação entre eles. A seguir nas Figuras 5 e 6 os diagramas de entidade e relacionamento montados com base nos nomes dos campos e conhecimento sobre seus significados.

Figura 5 - Diagrama Entidade e Relacionamento Base Cadastral Pessoa Jurídica



Fonte: Autoria própria. Usando Lucid Chart.

Figura 6 - Diagrama Entidade e Relacionamento Base Cadastral Pessoa Jurídica Simplificada



Fonte: Autoria própria. Usando Lucid Chart.

5.5. Planejamento

Pressman (p.15-16) elenca 5 atividades genéricas que podem ser aplicadas em vários tipos de desenvolvimento. São elas: Comunicação, Planejamento, Modelagem, Construção e Implantação. Além dessas atividades, ele também define 7 atividades guarda chuvas: Acompanhamento e controle de *software*, Gerenciamento de Risco, Controle de Qualidade de *Software*, Revisão Técnica, Medição, Gestão de Configuração, Gestão de Reusabilidade e Preparação e Produção do produto de trabalho. Algumas dessas atividades são difíceis de descrever e de pouca relevância para detalhar nesse documento. Como, por exemplo, "***Planejamento***", o desenvolvimento desse projeto foi feito somente por dois alunos de graduação autores do texto. A somente duas partes envolvidas e nenhuma hierarquia. Como são poucos agentes não precisa de grande controle e a comunicação é

simples. Além disso, apesar de o fluxo do processo não ser linear facilita a compreensão do mesmo a descrição de maneira linear no documento. Sendo assim, tendo um conhecimento inicial da base, seção anterior, começaremos pela atividade “comunicação” e a engenharia de requisitos.

Antes de qualquer atividade técnica, como criação de *software*, é essencial entender o usuário deste *software*. Isto é, comunicar-se com o usuário e outros “*stakeholders*” para melhor definir os requisitos e funcionalidades do projeto. A engenharia de requisitos engloba comunicação e modelagem (Pressman, p. 120). Como fazer a comunicação neste projeto “simulado” ? Karolita (2023) ressalta que a engenharia de requisitos demanda a comunicação com os stakeholders, mas que criar Personas é um método válido para lidar com essa demanda. Outro ponto importante é ressaltar que o trabalho visa a criação do *data warehouse* e não de um aplicativo ou “*Data mart”* completo. Ou seja, o usuário deve ser um cientista de dados ou pessoa de uma equipe que tenha capacidades técnicas para fazer consultas (*queries*) no *data warehouse*. Tendo isso em mente foram criadas 3 “Personas” para o desenvolvimento do restante do trabalho:

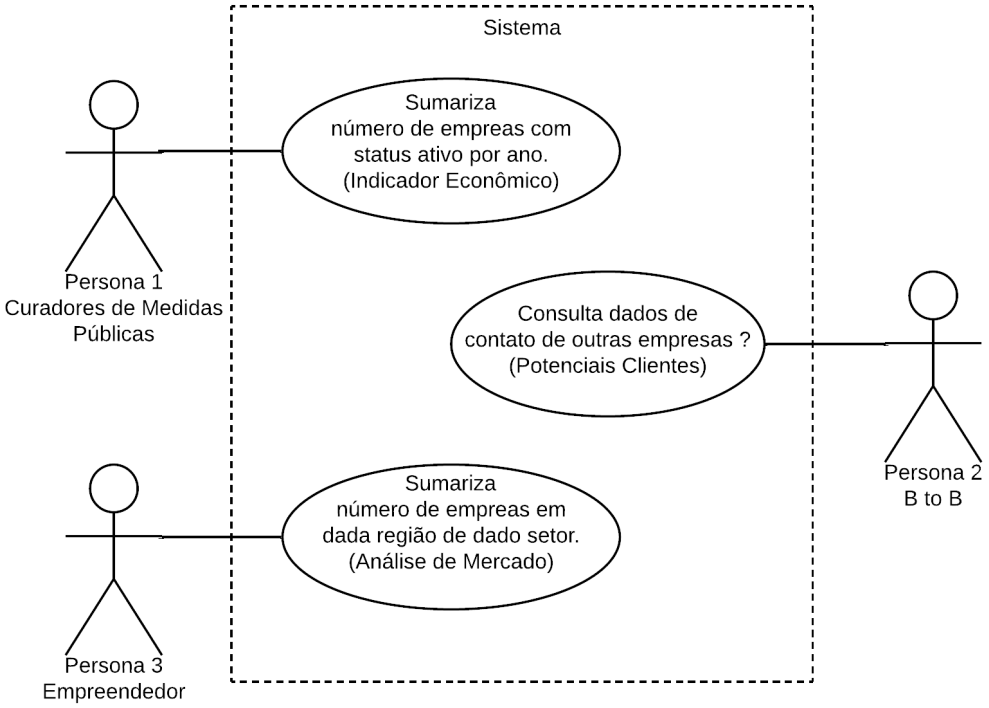
> ***Persona 1 - Curadores de medidas públicas:*** Seja um profissional do jornalismo ou parte de um grupo de cidadãos preocupados, essa Persona representa aqueles que querem consumir a base de cadastro de pessoas jurídicas com a finalidade de entender a situação macroeconômica atual, desempenho do governo, impacto de medidas econômicas e etc.
>
> ***Persona 2 - Empresário B to B :*** Esta Persona representa empresários e empresas que vêm a base como fonte de informação sobre o público alvo da sua empresa. Por exemplo, uma empresa de insumos alimentícios pode utilizar a base de cadastros de pessoas jurídicas para prospectar como clientes restaurantes da região onde atua.
>
> ***Persona 3 - Empreendedor :*** Este representa parte da iniciativa privada que pode utilizar a base para pesquisar oportunidades para abertura de empresa ou pesquisa de mercado. Não necessariamente empresas B to B, porém, empresários que queiram entender localização e nichos do mercado já atendidos.

A criação das “Personas” talvez não seja o ideal para o desenvolvimento do projeto. No entanto, supre as necessidades e limitações deste trabalho acadêmico. Entendemos que o principal motivo das bases públicas é a “Persona 1”, porém, achamos razoável a utilização da base nos outros contextos.

5.5.1. Diagrama de casos de Uso

As Personas e os casos de usos imaginados para o trabalho são de alto nível de abstração. Vão servir de guia para os requisitos do *Data Warehouse*, porém, vamos limitar o escopo do projeto somente até o *Data Warehouse*. Segue o diagrama de casos de uso, Figura 7, considerando as três personas:

Figura 7 - Diagrama De Casos de Uso



Fonte: Autoria própria. Usando Lucid Chart.

5.5.2. Requisitos Funcionais

Requisitos funcionais são os requisitos que descrevem o que o sistema deve fazer (Somerville, p.85). No Quadro 14 são elencados quatro requisitos funcionais do *data warehouse*.

Quadro 14 - Requisitos Funcionais

#	Requisito Funcionais	Motivação
01	O *Data Warehouse* deve estar atualizado.	Para consultas principalmente das PS2 e PS3 o dado deve ser tempestivo. Como a fonte de dados é uma fonte mensal, acreditamos que a atualização do sistema deva ser minimamente mensal. No entanto, sem muitos recursos é possível ter a atualização diária da base. Por outro lado, uma análise da informação utilizada pode demonstrar que o dado se altera pouco em poucos meses.
02	O *Data Warehouse* deve permitir consultar dados históricos.	Principalmente para PS1 alterações de indicadores decorrem da comparação com um baseline. Logo, o histórico se torna necessário.
03	O *Data Warehouse* não deve permitir duplicidade de dados.	Para consultas principalmente das PS2 e PS3 o dado deve ser unívoco. Por exemplo, não faz sentido um estabelecimento estar em dois lugares ao mesmo tempo.
04	Os dados obtidos através do *Data Warehouse* devem ser válidos.	Exemplo. Não faz sentido um número de telefone com 50 dígitos.

Fonte: Autoria própria.

5.5.3. Requisitos Não funcionais

Os requisitos não funcionais não apresentaram nenhuma especificidade e podem ser entendidos como os mesmos presentes no Quadro 02 seção 4.2.2.

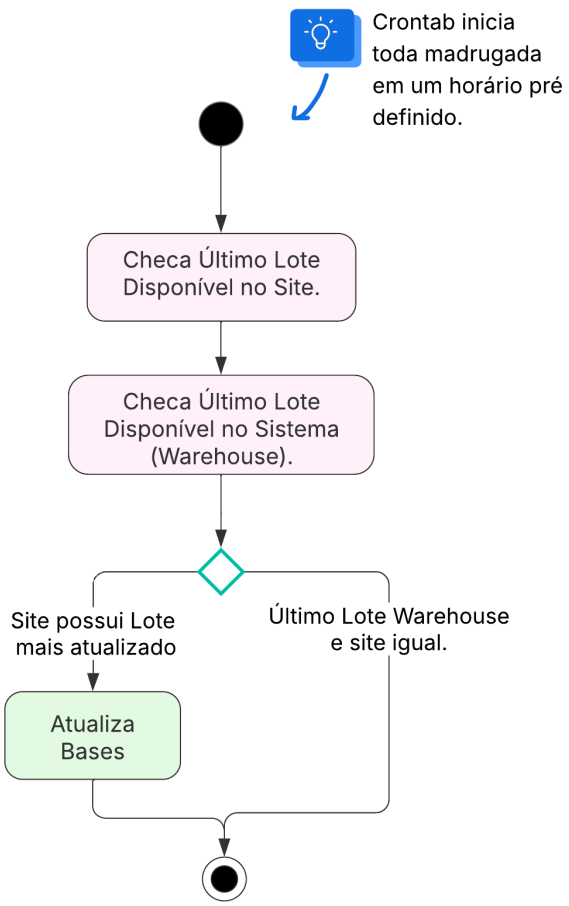
5.6. Construção efetiva dos do *data warehouse*

Após as primeiras etapas de entendimento e planejamento é necessária a construção efetiva do *data warehouse* seguindo os componentes da Figura 04. Logo, nas próximas subseções segue os componentes do *data warehouse* com algumas especificidades do estudo de caso em questão.

5.6.1. Gatilho

A fonte de informações do projeto é atualizada mensalmente. Sem um dia definido, variando entre dia 10 e 25 é disponibilizado no site do governo dentro de uma pasta atualizada com ANO-MES (YYYY-MM) os arquivos referente a atualização daquele mês. O gatilho implementado uma vez ao dia verifica o último lote (data de referência) disponível no site. Depois verifica o último lote salvo no sistema, caso eles sejam iguais, ele encerra o processo, caso tenha um novo lote disponível chama o subsistema de Atualização da Base (extração). Figura 8, representa o fluxo descrito acima.

Figura 8 - Fluxo do Gatilho de Atualização



Fonte: Autoria própria. Utilizando LucidChart.

5.6.2. Extração

Atualmente os arquivos são separados em: arquivos de empresas, 10 de estabelecimentos, 10 de sócios e outros 7 arquivos únicos. Dentro dos ZIP o nome dos arquivos CSVs não seguem um padrão de fácil leitura assim como o nome do ZIP, com isso na hora da extração renomeamos o CSV para o mesmo nome do ZIP em questão, com isso finalizado estamos prontos para a extração dos dados.

Os arquivos públicos têm seus dados separados por ***ponto-e-vírgula***. Contudo, ***ponto-e-vírgula*** não aparece somente como separador das colunas, mas também dentro dos conteúdos textuais das colunas dos campos. Este é um ponto de atenção para o tratamento e extração de qualquer base de dados em formato de texto. De preferência, o separador dos campos não deve ser um caracter que também esteja presente nos campos. Neste trabalho podemos contornar essa dificuldade entendendo como separador não apenas o ***ponto-e-vírgula***, mas ***fecha aspas ponto-e-vírgula abre aspas***. Essa sequência não usual, não faz parte dos textos de colunas texto. No entanto, ao utilizar esta sequência como separador tivemos que tratar a primeira e última coluna dos arquivos com um maior cuidado. Pois, a primeira coluna ficava com um ***abre aspas*** a mais antes do primeiro dado e a última coluna com um ***fecha aspas*** após o último dado da entrada correspondente. Outro ponto relevante de atenção são as variáveis do tipo data ou decimal. Pois, existem diversos formatos de data e normalmente os textos de decimais no formato brasileiro, “dia/mês/ano”, não são automaticamente nas funções de conversão (*casting*) do pyspark por padrão.

5.6.3. Limpeza & Padronização

A limpeza e padronização apesar de serem tratadas com funcionalidades diferentes ocorrem efetivamente ao mesmo tempo. Por exemplo, ao verificar que o corpo dos CNPJs têm 8 dígitos numéricos e retirando o registro em caso contrário. Ao mesmo tempo estamos definindo o formato numérico de até 8 dígitos e limpando os dados não conformes. Efetivamente poderíamos salvar este mesmo campo em diferentes tabelas como “texto” e como “numérico”. No entanto, faz parte da padronização utilizarmos o mesmo formato para se referir a esse campo em diversas tabelas. Dessa maneira, fica mais simples a utilização e cruzamento de tabelas. Durante a extração algumas colunas já podem ser formatadas na definição de esquema na leitura dos arquivos, no entanto, algumas colunas como os do tipo data são formatadas após a “extração”.

Segundo Kimball (2004):

> “Limpeza e conformidade são as principais etapas em que o sistema ETL agrega valor. As outras etapas de extração e entrega são obviamente necessárias, mas apenas movem e reformatam os dados. A limpeza e a conformidade, na verdade, alteram os dados e fornecem orientações sobre se os dados podem ser usados para os fins pretendidos.” (Kimball, 2004 p.113)

Isto é, podemos entender que é nessa parte que vamos agregar valor aos dados brutos extraídos do portal do governo. E, que as vamos alterar e melhorar a qualidade dos dados. Portanto, essa etapa contempla os seguintes tratamentos:

- Verificação da não nulidade das chaves.
- Verificação do número de caracteres dos componentes do CNPJ.
- Verificação do número de dígitos dos telefones.
- Verificação da Completude dos campos de telefone.
- Verificação dos dígitos verificadores de um CNPJ.
- Verificação da coerência dos CEP.

Além disso, entendemos que nessa etapa entra o aspecto de decidirmos retirar as colunas referentes ao fax. Como citado anteriormente na seção 4.2.1 fax é uma coluna com pouco preenchimento e se refere a uma tecnologia antiga. Logo, não vemos proveito em manter a coluna que provavelmente não será utilizada nos casos de uso previstos.

5.7. Métricas e Resultados

Com a finalidade de mensurar o resultado das atividades realizadas nesse projeto é necessário estabelecer algumas métricas para avaliação. Nesse sentido foram elencadas e métricas: **Qualidade dos dados, Consistência e Armazenamento.** Para melhor descrever as métricas e resultados obtidos foi organizado o Quadro 15 abaixo:

Quadro 15 - Sumário Métricas e Resultados

#	Métrica	Descrição Ajuste
01	Qualidade de Dados	Relacionado a Limpeza e Padronização, sabemos que os DDD, prefixos de telefones paras as diferentes regiões do Brasil, são compostos por números de 11 à 99. Além disso, sabemos que os números de telefones fixos têm até 8 dígitos enquanto telefones móveis têm 9 dígitos (ANATEL, 2010). Podemos notar uma volumetria

		considerável em uma amostra de 25 milhões, cerca de 1.6 milhões tinham algum “erro” no padrão dos contatos. Ou seja, 6% da base com número incompatível. Pensando em PS2 essa limpeza pode evitar tentativas de ligações desnecessárias.
02	Qualidade de Dados	O CNPJ completo composto de corpo, número da filial e dígito verificador é a chave para a base desses estabelecimentos. A chave primária efetivamente é só o corpo e a filial, pois o dígito verificador pode ser calculado através deles (RECEITA FEDERAL, 2024). Logo, uma verificação de qualidade foi o re-cálculo desse valor. Neste caso observamos 0 erros.
03	Consistência	Existe uma correlação entre CEP e UF. Em alguns casos CEP pode ser somente de uma rua, mas existe uma relação entre CEP e estado. Achamos 94 empresas com dados inconsistentes. A inconsistência pode ser trabalhada de várias formas neste trabalho entendemos que CEP tem prioridade.
04	Consistência	MEI vs Porte. Dos 15,8 milhões de empresas MEI 1.790 são não são classificadas como Microempresa ou Empresas de pequeno porte. Isto é um indicador de que o porte ou marcação de MEI da empresa pode estar errado.
05	Consistência	MEI vs Número de Sócios. Dos 15,8 milhões de empresas MEI 238 apresentam mais de 1 sócio. Ou seja, é inconsistente com o fato de MEI ser individual.
06	Consistência	MEI vs Filial. Dos 15,8 milhões de empresas MEI 3.549 apresentam mais de 1 filial. Isto não deveria ocorrer para empresas MEI.
07	Armazenamento	Com relação a padronização é possível notar que ao transformar os arquivos texto, com colunas “string”, do arquivo da receita para o formato “ORC” e com colunas com formatos adequados tivemos uma otimização do uso

		do espaço. A redução do espaço utilizado foi de aproximadamente 75% no espaço utilizado pelas tabelas de “estabelecimentos” e 66% nas tabelas de “empresas”.

Notamos que nem sempre as medidas de limpeza têm muito impacto. No caso, 02 a limpeza não teve nenhum efeito. E, no caso 03 o impacto é muito pequeno. Se por um lado o efeito da limpeza é “pequeno” , por outro isso é um indício da base ser boa. Se atualmente as métricas da base estão boas, manter o log dessa limpeza e observar caso algum desses números mude pode ser um indício do deterioramento da qualidade desses dados. Destaca-se o ganho da otimização do espaço com uma porcentagem de 75% para o arquivo com mais informações. Além dessas métricas quantitativas, outros benefícios podem ser considerados ao se usar o *data warehouse*. O cientista de dados, ou outro usuário, não está necessariamente interessado em saber onde os arquivos estão disponíveis na rede e a configuração sugerida deixa o acesso ser realizado através do nome simplificado da tabela. Outro código importante é o arquivo que configura o “crontab” para atualização automática do arquivo, deixando os dados disponibilizados de maneira tempestiva.

6. CONCLUSÃO

No desenvolvimento da criação do *data warehouse* a partir da base pública de cadastros de pessoas jurídicas conseguimos avaliar a abordagem proposta. Essa abordagem e sua aplicação contempla conceitos de Engenharia de *Software*. Outros conceitos de engenharia de *software* relacionados a manutenção do código e controle de equipe não se aplicam neste projeto acadêmico. Pois, o sistema foi feito em duas pessoas com duração de um semestre. Logo, conseguimos observar que com uma uma validação e exploração inicial dos dados é essencial para posterior entendimento de relacionamento dos mesmo, o casos de uso dessas informações e requisitos, para só então a criação do *data warehouse*. Descrevemos as etapas do *data warehouse* de extração até a disponibilização dos dados. Já com o código criado para o *data warehouse* e com os requisitos de qualidade conseguimos validar algumas métricas de qualidade de dados, consistência e armazenamento. Conseguimos dessa maneira avaliar que cerca de 6% das entradas tem um número errado de contato, que pelo menos 3 mil empresas MEI tem mais que uma filial e que utilizando o formato correto conseguimos armazenar os dados com até 75% mais de eficiência.

6.1. Trabalhos Futuros

Para trabalhos futuros poderiam ser trabalhados outros ajustes possíveis das informações, mas também é possível a criação de data marts que consumam direto do *data warehouse* construído neste trabalho. Não exploramos neste trabalho a limpeza da base utilizando os dados históricos. Um exemplo de análise seria analisar a mudança da situação cadastral sabemos que a situação cadastral baixada não deveria voltar a ser ativa. Outra análise que ficou em aberto seria de empresas com o CNAE principal igual a 8888888, e como elas estão distribuídas entre as diferentes situações cadastrais. Uma terceira possibilidade é verificar as volumetrias demonstradas nas Análises Exploratórias usando somente empresas com situação cadastral Ativa que são mais relevantes. Já em um trabalho que tivesse o intuito de criar um data mart seria interessante demonstrar as informações que podemos tirar dos dados crus do *data warehouse*.

REFERÊNCIAS

ANATEL. **NUMERAÇÃO.** 17 nov. 2020. Disponível em: <https://www.gov.br/anatel/pt-br/regulado/numeracao>. Acesso em: 30 jun. 2025.

BRASIL. **Lei nº 12.527, de 18 de novembro de 2011.** Diário Oficial da União, Brasília, DF, 18 nov. 2011. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Acesso em: 09 mar. 2025.

https://www.gov.br/anatel/pt-br/regulado/numeracao

BRASIL. **Portal de Dados Abertos. Cadastro Nacional de Pessoa Jurídica - CNPJ.** Disponível em: <https://dados.gov.br/dados/conjuntos-dados/cadastro-nacional-da-pessoa-juridica---cnpj>. Acesso em: 09 mar. 2025.

BRASIL. **Mapa de Empresas - Empresas e Negócios.** Boletim do Mapa de Empresas. Disponível em: <https://www.gov.br/empresas-e-negocios/pt-br/mapa-de-empresas/>. Acesso em: 09 mar. 2025.

BRASIL. **Ministério do Empreendedorismo, da Microempresa e da Empresa de Pequeno Porte. Mapa de Empresas: Boletim do 3º quadrimestre de 2024.** 24 jan. 2025. Disponível em: <https://www.gov.br/empresas-e-negocios/pt-br/mapa-de-empresas/boletins/boletim-do-mapa-de-empresas-3o-quad-2024.pdf>. Acesso em: 09 mar. 2025.

BRASIL. **Ciência de dados - Governo Digital.** Disponível em: <https://www.gov.br/governodigital/pt-br/capacitacao/capacita-gov-br/ciencia-de-dados>. Acesso em: 03 jun. 2025.

CORREIOS. **Busca CEP.** Disponível em: <https://buscacepinter.correios.com.br/app/faixa_cep_uf_localidade/index.php>. Acesso em: 03 jun. 2025.

DONOHO, D. **50 Years of Data Science.** Journal of Computational and Graphical Statistics 26, 745-766, 2017. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/10618600.2017.1384734>. Acesso em: 16 mar. 2025.

ENAP. **Ciência de Dados em Políticas Públicas: uma experiência de formação.** Brasília, DF, 2022. Disponível em: <https://repositorio.enap.gov.br/bitstream/1/7472/2/Livro%20Digital%20Ci%C3%AAncia%20de%20Dados%20em%20Pol%C3%ADticas%20P%C3%BAblicas_compressed.pdf>. Acesso em: 16 mar. 2025.

FERREIRA, J.; ABELHA, A.; MACHADO, J. **O Processo ETL em Sistemas Data Warehouse.** II Simpósio de Informática. Braga, Portugal, 2010. Disponível em: <https://www.researchgate.net/profile/Jose-Machado-23/publication/265195317_O_Processo_ETL_em_Sistemas_Data_Warehouse/links/5580380a08aea3d7096e442e/O-Processo-ETL-em-Sistemas-Data-Warehouse.pdf>. Acesso em: 03/05/2025.

Google Trends. **Consulta a "Ciência de Dados" e "Engenharia de dados" no Brasil nos últimos 5 anos.** Disponível em: <https://trends.google.com/trends/explore?date=today%205-y&geo=BR&q=Ci%C3%AAncia%20de%20Dados,Engenharia%20de%20dados&hl=en>. Acesso em: 16 mar. 2025.

KAROLITA, D.; MCINTOSH, J.; KANIJ, T.; GRUNDY, J.; OBIE, H. **Use of Personas in Requirements Engineering: A systematic mapping study.**Information and Software Technology, v. 162, October 2023.

KIMBALL, Ralph; CASERTA, Joe. **The Data Warehouse ETL Toolkit:** Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data. Editora Wiley, 2004.

MACHADO, A. **Construção de um Processo ETL Automatizado em Dados de Campanhas de uma Empresa no Setor Bancário.** Trabalho de Conclusão de Curso - Faculdade de Gestão de Negócios, Universidade Federal de Uberlândia. Uberlândia, MG, 2023. Disponível em: <https://repositorio.ufu.br/bitstream/123456789/38346/1/Construc%c3%a7%c3%a3oProcessoETL.pdf>. Acesso em: 29 mar. 2025.

PARRACHO, T. M.; ZACARIAS, R. O.; SANTOS, R. P. **Os efeitos da experiência de desenvolver no processo de ensino-aprendizagem de engenharia de software.** EDT - Educação Temática Digital. Campinas, SP, 2025. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/etd/article/view/8674823/35849>. Acessado em: 29 mar. 2025.

PRESSMAN, R. S. **Engenharia de Software**. 7a. ed. McGraw-Hill. 2010.

RAUTENBERG, S.; CARMO, P. R. V. **Big data e ciência de dados: complementaridade conceitual no processo de tomada de decisão.** Brazilian Journal of Information Science, v. 13, n., 2019. Disponível em: <https://dialnet.unirioja.es/descarga/articulo/6983493.pdf>. Acesso em: 11 mar. 2025.

RECEITA FEDERAL. **MANUAL DE CÁLCULO DO DV DO CNPJ.** 05 nov. 2024. Disponível em: <https://www.gov.br/receitafederal/pt-br/centrais-de-conteudo/publicacoes/documentos -tecnicos/ cnpj/manual-dv-cnpj.pdf/view>. Acesso em: 30 jun. 2025.

RODRIGUES, João Gaspar. **Publicidade, transparência e abertura na administração pública.** Revista de Direito Administrativo (RDA), Rio de Janeiro, v. 266, p.89-123, mai/ ago 2014. Disponível em:<https://periodicos.fgv.br/rda/article/view/32142/30937>. Acesso em: 10 mar. 2025.

SEMIDÃO, Rafael Aparecido Moron. **Dados, informação e conhecimento enquanto elementos de compreensão do universo conceitual da ciência da informação: contribuições teóricas.** 2014. 198 f. Tese (Mestrado) - Curso de Ciência da Informação, Universidade Estadual Paulista, Marília, 2014. Disponível em: <https://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/semidao_ram_me_mar.pdf>. Acesso em: 10 mar. 2025.

SOMMERVILLE, Ian. **Engenharia de Software**. 10a. ed. Editora Pearson, 2011.

ANEXO I

Neste anexo estão presentes os dados brutos da tabela pública. 5 primeiras linhas de cada arquivo.

Arquivo CNAES 5 primeiras linhas.

"0111301";"Cultivo de arroz"
"0111302";"Cultivo de milho"
"0111303";"Cultivo de trigo"
"0111399";"Cultivo de outros cereais não especificados anteriormente"
"0112101";"Cultivo de algodão herbáceo"

Arquivo Motivos 5 primeiras linhas.

"00";"SEM MOTIVO"
"01";"EXTINCAO POR ENCERRAMENTO LIQUIDACAO VOLUNTARIA"
"02";"INCORPORACAO"
"03";"FUSAO"
"04";"CISAO TOTAL"
"05";"ENCERRAMENTO DA FALENCIA"

Arquivo Munícipio 5 primeiras linhas.

"0001";"GUAJARA-MIRIM"
"0002";"ALTO ALEGRE DOS PARECIS"
"0003";"PORTO VELHO"
"0004";"BURITIS"
"0005";"JI-PARANA"

Arquivo Países 5 primeiras linhas.

"000";"COLIS POSTAUX"
"013";"AFEGANISTAO"
"017";"ALBANIA"
"020";"ALBORAN-PEREJIL,ILHAS"
"023";"ALEMANHA"

Arquivo Qualificações 5 primeiras linhas:

"00";"Não informada"
"05";"Administrador"
"08";"Conselheiro de Administração"
"09";"Curador"
"10";"Diretor"

Arquivo Natureza 5 primeiras linhas:

"0000";"Natureza Jurídica não informada"
"3271";"Órgão de Direção Local de Partido Político"
"3280";"Comitê Financeiro de Partido Político"
"3298";"Frente Plebiscitária ou Referendária"
"3301";"Organização Social (OS)"

Arquivo Empresas 5 primeiras linhas:
Usamos o símbolo * para omitir a informação mesmo o dado estando aberto publicamente.

"4******2";"4**********A";"2135";"50";"30000,00";"01";""
"4******3";"J**********5";"50";"3000,00";"01";""
"4******4";"O**********87";"2135";"50";"5000,00";"01";""
"4******5";"G**********0";"2135";"50";"3000,00";"01";""
"4******6";"R**********1";"2135";"50";"10000,00";"01";""

Arquivo Sócios 5 primeiras linhas:

"4******4";"2";"M**********S";"***386625**";"49";"19911204";"";"***000000**";"";"00";"6"
"4******4";"2";"A**********S";"***243095**";"22";"20150529";"";"***000000**";"";"00";"6"
"0******8";"2";"L**********O";"***272303**";"49";"20020207";"";"***000000**";"";"00";"8"
"0******8";"2";"P**********O";"***557645**";"49";"20020207";"";"***000000**";"";"00";"5"
"0******3";"1";"B**********.";"0**********5";"22";"20010216";"";"***000000**";"";"00";"0"

Arquivo Estabelecimento 5 primeiras linhas:

linha 1 :

"3******9";"0001";"95";"1";"";"02";"20190711";"00";"";"";"20190711";"4771701";"";"RUA";"TAMOIO";"1233";"";"NITEROI";"92120001";"RS";"8589";"51";"95250775";"";"";"";"";"";"";""

linha 2 :

"3******4";"0001";"33";"1";"A******O";"08";"20200612";"01";"";"";"20190711";"5612100";"";"RUA";"DOUTOR DEVALDO BORGES";"185";"APT 22";"JARDIM SAO PAULO";"50910390";"PE";"2531";"81";"9******48";"";"";"";"";"";"A******2@GMAIL.COM.BR";"";""

linha 3:

"3******7";"0001";"02";"1";"";"04";"20220107";"63";"";"";"20190711";"4330404";"";"RUA";"ROCHA POMBO";"926";"";"CASCAVEL";"85025020";"PR";"7583";"42";"9******9";"";"";"";"";"";"";""

linha 4:

"3******9";"0001";"37";"1";"";"04";"20220112";"63";"";"";"20190711";"4399103";"";"RUA";"DOIS";"57";"";"CONFORTO";"27265435";"RJ";"5925";"71";"9******0";"";"";"";"";"";"r******s@hotmail.com";"";""

linha 5:

"3******0";"0001";"05";"1";"";"04";"20221123";"63";"";"";"20190711";"9602501";"";"PASSAGEM";"CABRAL";"51";"CONJ JARDIM DAS ORQUMDEAS";"ICUI-GUAJARA";"67125210";"PA";"0415";"91";"8******0";"";"";"";"";"";"a******7@gmail.com";"";""

Arquivo Simples 5 primeiras linhas:

"0******0";"N";"20070701";"20070701";"N";"20090701";"20090701"

"0******6";"N";"20180101";"20191231";"N";"00000000";"00000000"

"0******8";"N";"20140101";"20211231";"N";"00000000";"00000000"

"0******1";"S";"20070701";"00000000";"N";"00000000";"00000000"

"0******3";"N";"20090101";"20231231";"N";"00000000";"00000000"