UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
SISTEMAS DE INFORMAÇÃO

LUIS FELIPE ARAUJO MOTA

Um Estudo Comparativo de Ferramentas para Perfilamento de Dados em Larga Escala

Recife

2025

UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
SISTEMAS DE INFORMAÇÃO

LUIS FELIPE ARAUJO MOTA

Um Estudo Comparativo de Ferramentas para Perfilamento de Dados em Larga Escala

Trabalho de Conclusão de Curso apresentado no curso de Bacharelado em Sistemas de Informação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador(a): ROBSON FIDALGO

Recife

2025

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Mota, Luis Felipe Araujo.
Um estudo comparativo de ferramentas para perfilamento de dados em larga escala / Luis Felipe Araujo Mota. - Recife, 2025.
55 p. : il., tab.

Orientador(a): Robson Fidalgo
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Sistemas de Informação - Bacharelado, 2025.
Inclui referências, anexos.

1. Big data. 2. Perfilamento de dados. 3. Qualidade de dados. 4. Ferramentas de análise. 5. Engenharia de dados. 6. Avaliação da qualidade de dados. I. Fidalgo, Robson. (Orientação). II. Título.

000 CDD (22.ed.)

LUIS FELIPE ARAUJO MOTA

Um Estudo Comparativo de Ferramentas para Perfilamento de Dados em Larga Escala

Trabalho de Conclusão de Curso apresentado ao Centro de Informática (CIn) da Universidade Federal de Pernambuco (UFPE), como requisito parcial para obtenção do título de bacharel em Sistemas de Informação.

.

Aprovado em: 22/08/2025.

BANCA EXAMINADORA

_____

Prof. Dr. Robson do Nascimento Fidalgo (Orientador)
Universidade Federal de Pernambuco

_____

Prof. Dr. Vinícius Cardoso Garcia (Examinador Interno)
Universidade Federal de Pernambuco

AGRADECIMENTOS

Agradeço primeiramente a Deus, por ter me concedido forças nos momentos de dificuldade, sabedoria para seguir em frente e fé para nunca desistir dos meus objetivos. Sua presença constante foi essencial para que eu superasse os desafios ao longo dessa caminhada acadêmica. Sem ele, nada disso teria sido possível.

À minha família, meu mais profundo agradecimento, especialmente à minha mãe, Carla, que sempre foi o maior exemplo de dedicação aos estudos e à pesquisa. Seu amor incondicional, apoio firme e inspiração constante foram fundamentais para a minha formação, tanto acadêmica quanto pessoal. Agradeço também ao meu pai, Pedro, e ao meu irmão, Michel, por me ensinarem que o amor verdadeiro também corrige e fortalece. Sem vocês, essa conquista não teria o mesmo sentido.

Aos professores e colegas que cruzaram meu caminho no CIn, deixo minha sincera gratidão. Agradeço também aos amigos que se tornaram parte da minha família escolhida, cuja parceria, incentivo e companheirismo tornaram minha experiência universitária mais leve, rica e significativa.

“A conquista é um acaso que talvez dependa mais das falhas dos vencidos do que do gênio do vencedor.”

Madame de Staël

RESUMO

O crescimento exponencial na geração de dados, impulsionado por sistemas digitais, sensores e plataformas em rede, tem transformado o cenário da engenharia de dados, especialmente com o advento do paradigma Big Data. Nesse contexto, a compreensão e a qualidade dos dados assumem papel estratégico para organizações que buscam decisões fundamentadas em evidências confiáveis. O data profiling, entendido como o processo sistemático de extração de metadados estatísticos e estruturais, emerge como etapa crítica para a inspeção, limpeza e integração de dados, sobretudo em ambientes caracterizados por grande volume e diversidade de formatos. No entanto, a aplicação prática do data profiling em cenários de Big Data ainda carece de estudos que combinem rigor técnico e análise funcional. Este trabalho apresenta um estudo funcional e comparativo de três ferramentas de data profiling com suporte a ambientes de dados em larga escala. A partir de critérios metodológicos de seleção e de um checklist funcional baseado em literatura especializada, as ferramentas são avaliadas quanto às suas funcionalidades e desempenho frente a conjuntos de dados públicos representativos. Os resultados obtidos permitem identificar as vantagens, limitações e melhores contextos de uso de cada solução, contribuindo para a escolha fundamentada de ferramentas de data profiling em projetos de engenharia de dados.

Palavras-chave: Big data. Perfilamento de dados. Qualidade de dados. Ferramentas de análise. Engenharia de dados. Avaliação da qualidade de dados

ABSTRACT

The exponential growth in data generation, driven by digital systems, sensors and networked platforms, has reshaped the data engineering landscape, especially with the rise of the Big Data paradigm. In this scenario, data understanding and quality play a strategic role for organizations seeking decisions based on reliable evidence. Data profiling, understood as the systematic process of extracting statistical and structural metadata, emerges as a critical step for data inspection, cleansing and integration, especially in environments characterized by high volume and format diversity. However, the practical application of data profiling in Big Data scenarios still lacks studies that combine technical rigor with functional analysis. This work presents a functional and comparative study of three data profiling tools designed for large-scale data environments. Based on methodological selection criteria and a functional checklist grounded in specialized literature, the tools are evaluated regarding their capabilities and performance against representative public datasets. The results allow the identification of strengths, limitations, and best use cases for each solution, contributing to informed decision-making in data profiling tool selection for data engineering projects.

Keywords: Big Data. Data profiling. Data quality. Analysis tools. Data engineering. DQ Evaluation

LISTA DE ILUSTRAÇÕES

Figura 1 –	Características do big data	17
Figura 2 –	Ciclo do Big Data	18
Figura 3 –	Tipos de armazenamento do Big Data	19
Figura 4 –	Taxonomia clássica das tarefas de Data Profiling	23
Figura 5 –	Catálogo de requisitos de Perfilamento de dados de ferramentas de Qualidade de Dados	30

LISTA DE TABELAS

Tabela 1 –	Principais tipos de tecnologias integradas em ambientes Lakehouse	20
Tabela 2 –	Critérios de Inclusão Utilizados na Seleção das Ferramentas	27
Tabela 3 –	Plataformas de pesquisa de Ferramentas de Perfilamento de Dados	28
Tabela 4 –	Checklist de Tarefas de Perfilamento de dados	30
Tabela 5 –	Legenda Para o Checklist	36
Tabela 6 –	Checklist de Tarefas de Perfilamento de dados aplicada às três ferramentas	36
Tabela 7 -	Critérios qualitativos de adoção e integração das ferramentas analisadas	40
Tabela 8 –	Características dos datasets utilizados para teste de desempenho	44
Tabela 9 –	Resultados do teste de desempenho	45

LISTA DE ABREVIATURAS E SIGLAS

HDFS	Sistema de Arquivos Distribuído Hadoop
DQ Evaluation	Verificação da Qualidade de Dados
3 V’s	Características do Big Data (Velocidade, Variedade e Volume)
ACID	Atomicidade, Consistência, Isolamento e Durabilidade
OLAP	Processamento Analítico Online
DQ	Data Quality (Qualidade de Dados)
SDK	Kit de Desenvolvimento de Software
API	Interface de Programação de Aplicações

SUMÁRIO

1	**INTRODUÇÃO**	13
1.1	CONTEXTUALIZAÇÃO	13
1.2	MOTIVAÇÃO	14
1.3	OBJETIVOS	15
1.4	ESTRUTURA DO TRABALHO	15
2	**REFERENCIAL TEÓRICO**	17
2.1	BIG DATA: FUGINDO DA COMPUTAÇÃO TRADICIONAL	17
2.2	PERFILAMENTO DE DADOS	21
2.3	TAXONOMIA DE TAREFAS DE PERFILAMENTO DE DADOS	22
2.3.1	PERFILAMENTO DE COLUNA ÚNICA	24
2.3.2	PERFILAMENTO DE MÚLTIPLAS COLUNAS	24
2.3.3	DESCOBERTA DE DEPENDÊNCIAS	25
3	**FERRAMENTAS PARA PERFILAMENTO DE DADOS EM LARGA ESCALA**	27
3.1	PROCESSO DE SELEÇÃO DAS FERRAMENTAS	27
3.2	MÉTODO DE MENSURAÇÃO FUNCIONAL DAS FERRAMENTAS	28
3.3	ANÁLISE FUNCIONAL	32
3.3.1	DATABRICKS	32
3.3.2	YDATAPROFILING	34
3.3.3	DATAEDO	35
3.3.4	TABELA COMPARATIVA DAS FERRAMENTAS	36
3.4	CRITÉRIOS QUALITATIVOS DE ADOÇÃO E INTEGRAÇÃO	39
3.5	FERRAMENTAS EMERGENTES NÃO AVALIADAS	41
4	**ANÁLISE DE DESEMPENHO DE FERRAMENTAS PARA PERFILAMENTO DE DADOS EM LARGA ESCALA**	42
4.1	ESCOLHA DOS CONJUNTOS DE DADOS	43
4.2	EXECUÇÃO DAS FERRAMENTAS	44
4.3	LIMITAÇÕES DO CONJUNTO DE TESTES PRÁTICOS	46
5	**CONSIDERAÇÕES FINAIS**	48
6	**TRABALHOS FUTUROS**	49
	REFERÊNCIAS	50

ANEXO A – EVIDÊNCIAS VISUAIS DA FERRAMENTA DE PERFILAMENTO DE DADOS DO DATABRICKS 53
ANEXO B – EVIDÊNCIAS VISUAIS DA FERRAMENTA DE PERFILAMENTO DE DADOS DO YDATAPROFILING 54
ANEXO C – EVIDÊNCIAS VISUAIS DA FERRAMENTA DE PERFILAMENTO DE DADOS DO DATAEDO 56

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

A sociedade contemporânea está profundamente marcada pela intensificação do uso de tecnologias digitais e pela consequente explosão na geração de dados em múltiplos domínios. Transações financeiras, interações sociais, dispositivos móveis, sensores ambientais, plataformas de e-commerce e sistemas corporativos produzem continuamente grandes quantidades de dados em formatos variados, dentre eles os estruturados, semiestruturados e não estruturados. Segundo Taleb, Serhani e Dssouli (2019), esse fenômeno impulsionou o surgimento do termo Big Data, que ultrapassa a simples noção de grande volume e passa a incorporar múltiplas dimensões, como a velocidade de geração, a variedade de formatos, a veracidade das informações, entre outros aspectos que influenciam diretamente sua manipulação e aproveitamento.

Com o aumento da complexidade e da escala dos dados disponíveis, surgem novos desafios relacionados à sua coleta, armazenamento, processamento, análise e governança. O paradigma do Big Data impõe mudanças significativas nos métodos tradicionais de tratamento de dados, exigindo abordagens distribuídas, escaláveis e adaptáveis ao dinamismo das fontes de informação. Dai et al. (2016) destacam que o verdadeiro valor do dado não está apenas em sua existência, mas na capacidade de transformá-lo em conhecimento acionável, por meio de processos confiáveis que garantam sua qualidade, integridade e contextualização. Nesse contexto, o conceito de qualidade de dados ganha centralidade, sendo compreendido como o grau em que um conjunto de dados atende às necessidades de uso em termos de completude, consistência, precisão, atualidade e confiabilidade.

Ainda que o avanço de técnicas de análise como o data mining tenha ampliado a capacidade de gerar insights e padrões a partir de grandes volumes de dados, esses processos analíticos dependem fortemente da condição dos dados de entrada. Dados ruidosos, incompletos ou ambíguos tendem a comprometer a eficácia dos resultados. Por isso, etapas como o data cleansing, responsável por identificar e corrigir erros e inconsistências, e o data profiling, voltado à inspeção detalhada da estrutura e do conteúdo dos dados, tornaram-se etapas fundamentais em fluxos de trabalho baseados em dados.

Compreender os dados, portanto, é uma exigência elementar na era digital, especialmente em ambientes onde a informação se apresenta de maneira desorganizada,

incompleta ou não documentada. O Big Data, nesse sentido, não representa apenas uma oportunidade tecnológica, mas também uma demanda por técnicas refinadas de diagnóstico e entendimento que sustentem o uso estratégico dos dados em diversas áreas da sociedade. A análise antecipada e consciente dos dados tornou-se parte inseparável de qualquer processo robusto de tomada de decisão baseada em informação.

1.2 MOTIVAÇÃO

A consolidação do Big Data como pilar estratégico em diferentes setores da sociedade trouxe consigo não apenas novas oportunidades tecnológicas, mas também desafios complexos no que se refere à gestão, qualidade e compreensão dos dados. Em ambientes caracterizados por grande volume, variedade e velocidade de informação, torna-se imprescindível adotar práticas capazes de diagnosticar a estrutura e os padrões dos dados antes de seu uso analítico. A ausência de conhecimento prévio sobre o conteúdo e a organização das bases de dados pode comprometer seriamente a eficiência de modelos de análise, a confiabilidade de decisões automatizadas e a integridade de sistemas de informação.

Nesse contexto, o perfilamento de dados surge como uma etapa fundamental. Trata-se de um processo que visa inspecionar e descrever os dados de maneira exploratória, gerando metadados que facilitam a compreensão de sua estrutura, distribuição, qualidade e possíveis inconsistências. Conforme apontam Abedjan, Golab e Naumann (2015), o data profiling é essencial em fluxos de trabalho que envolvem integração, limpeza e preparação de dados, atuando como diagnóstico inicial em projetos analíticos e operacionais.

Apesar de sua reconhecida relevância teórica, observa-se uma carência significativa de estudos práticos e aprofundados que investiguem o uso do data profiling em cenários de Big Data. A maior parte da literatura concentra-se em ambientes tradicionais, como bancos relacionais ou datasets de pequena escala, o que pouco reflete os desafios reais enfrentados por profissionais que atuam com dados massivos e heterogêneos. Há, sobretudo, uma lacuna no que se refere à análise funcional de ferramentas existentes para o perfilamento e à realização de testes práticos em ambientes estruturados para volumes massivos de dados.

Essa ausência de referenciais claros dificulta não apenas a adoção estratégica de soluções de data profiling, mas também a construção de metodologias robustas que possam orientar sua aplicação em contextos empresariais ou científicos. Diante disso, a motivação central deste trabalho reside na necessidade de suprir essa lacuna, por meio de um estudo que alie fundamentação conceitual sólida à investigação aplicada, voltada à compreensão

funcional do perfilamento de dados em ambientes de Big Data, com atenção especial à categorização de tarefas, limitações técnicas e critérios objetivos de avaliação.

1.3 OBJETIVOS

Este trabalho tem como objetivo principal realizar um estudo funcional de ferramentas de data profiling aplicadas a ambientes de Big Data, analisando sua capacidade de lidar com desafios específicos como volume massivo de dados, diversidade de formatos e ausência de esquemas estruturados. Ao investigar o comportamento dessas ferramentas em cenários realistas, busca-se compreender em que medida elas conseguem atender às necessidades práticas de profissionais e organizações que operam com grandes volumes de dados.

Especificamente, o estudo propõe a avaliação funcional comparativa de três ferramentas selecionadas com base em critérios de relevância técnica, disponibilidade gratuita e compatibilidade com formatos amplamente utilizados em ambientes distribuídos, como o Delta Lake. A análise é conduzida a partir de um checklist funcional fundamentado em literatura especializada e adaptado a práticas contemporâneas de data profiling, com foco em caracterização de dados, verificação de qualidade e identificação de padrões estruturais. Adicionalmente, são realizados experimentos com conjuntos de dados públicos, de modo a testar o desempenho prático das ferramentas em tarefas comuns de inspeção e diagnóstico de dados em larga escala.

Ao final, pretende-se oferecer um panorama claro das funcionalidades disponíveis, das limitações enfrentadas por cada ferramenta e de suas respectivas adequações para distintos cenários de aplicação, contribuindo com subsídios úteis para decisões técnicas no campo da engenharia de dados.

1.4 ESTRUTURA DO TRABALHO

Este trabalho está organizado em cinco capítulos, além das seções de referências e apêndices. O Capítulo 1 apresenta a introdução ao tema, abordando a contextualização, a motivação que impulsionou a pesquisa, os objetivos a serem alcançados e a organização geral do documento. O Capítulo 2 trata do referencial teórico necessário à compreensão do escopo do trabalho, discutindo os fundamentos do Big Data e os principais conceitos relacionados ao data profiling, bem como suas classificações técnicas e implicações em contextos de dados em larga escala.

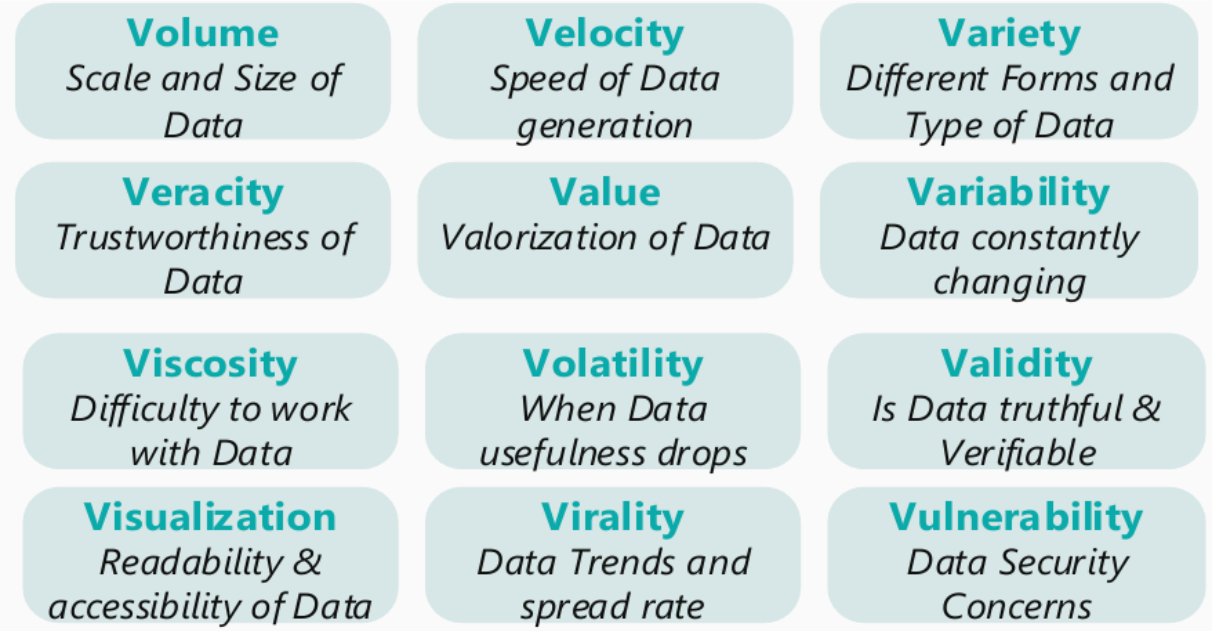
O Capítulo 3 descreve as ferramentas de data profiling selecionadas, justificando os critérios utilizados na escolha e detalhando a metodologia aplicada para avaliação funcional. Em seguida, o Capítulo 4 apresenta a análise de desempenho das ferramentas em cenários simulados de Big Data, com base na execução prática de tarefas e na observação de métricas relevantes. O Capítulo 5 expõe as considerações finais, sintetizando os principais resultados obtidos, apontando as limitações da pesquisa . Por fim, o Capitulo 6 sugere possíveis direções para trabalhos futuros.

2 REFERENCIAL TEÓRICO

2.1 BIG DATA: FUGINDO DA COMPUTAÇÃO TRADICIONAL

O termo Big Data emergiu no início dos anos 2000 como uma resposta à crescente incapacidade de sistemas tradicionais de processamento lidarem com o enorme volume de dados gerados por dispositivos digitais, redes sociais, sensores e aplicações empresariais. Segundo Taleb, Serhani e Dssouli (2019), o conceito representa um novo paradigma tecnológico que exige arquiteturas e ferramentas especializadas para armazenamento, gerenciamento e análise de dados em escala massiva

Figura 1 – Características do big data

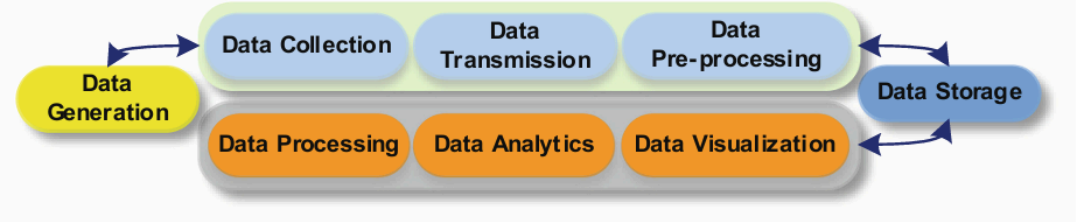


Fonte: Taleb et al., 2019

Segundo Taleb, Serhani e Dssouli (2019), inicialmente o Big Data era definido pelos chamados “3Vs”: Volume, Velocidade e Variedade. O volume refere-se à quantidade massiva de dados gerados continuamente; a velocidade diz respeito à rapidez com que esses dados são produzidos e precisam ser processados; e a variedade corresponde à diversidade de formatos, fontes e estruturas — como dados estruturados, semiestruturados e não estruturados. Com o tempo, essa definição foi expandida para até 12Vs, incluindo atributos como veracidade, valor, validade, variabilidade, visualização, volatilidade, entre outros, como representado na Figura 1 do trabalho.

A crescente complexidade desses dados impõe desafios significativos à escalabilidade dos sistemas de informação. Abedjan, Golab e Naumann (2015) destacam que muitas das tarefas tradicionais de análise de dados — como ordenação, agrupamento e correlação — tornam-se computacionalmente inviáveis quando aplicadas a grandes volumes de registros e atributos. A escalabilidade, portanto, não é apenas uma questão de velocidade de processamento, mas envolve a necessidade de arquiteturas distribuídas, particionamento eficiente, paralelismo e técnicas de sampling ou lazy evaluation para garantir desempenho aceitável.

Figura 2 – Ciclo do Big Data

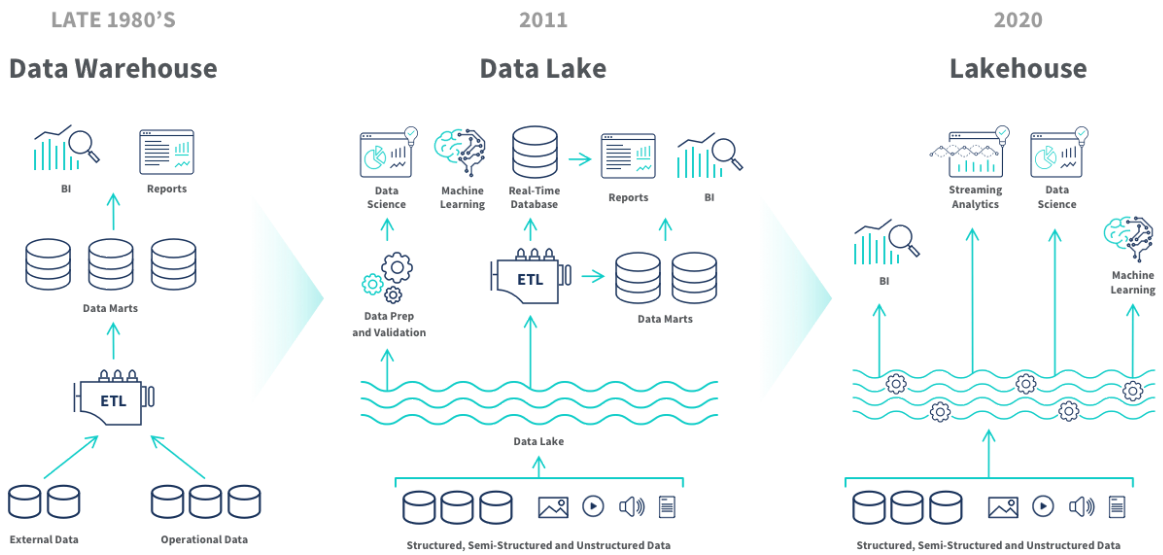


Fonte: Taleb et al., 2019

Segundo Dai et al. (2016), esses desafios técnicos estão intimamente ligados ao ciclo do Big Data, que compreende as etapas de geração, coleta, transmissão, armazenamento, processamento, análise e visualização dos dados. A geração pode ocorrer por meio de sensores, logs, APIs, transações e mídias sociais. Já a coleta e a transmissão requerem protocolos escaláveis e sistemas distribuídos de ingestão, como o Apache Kafka e o Flume. O armazenamento geralmente é feito em data lakes ou lakehouses, enquanto a análise é realizada por motores como Apache Spark, Dask ou Hadoop.

Nesse contexto, diversos modelos de armazenamento são utilizados para atender às exigências de escalabilidade e flexibilidade.

Figura 3 – Tipos de armazenamento do Big Data



Fonte: Insight software, 2023

Segundo Schneider et al. (2023) e Mazumdar et al. (2023), o Data Warehouse é uma plataforma de dados voltada para análise empresarial e suporte à decisão. Sua arquitetura é fundamentada em modelos relacionais com esquemas rígidos, definidos no momento da escrita, sendo ideal para consultas OLAP (Online Analytical Processing) e geração de relatórios históricos. Esses sistemas garantem propriedades ACID (Atomicidade, Consistência, Isolamento e Durabilidade), além de oferecer recursos avançados de governança, versionamento e desempenho otimizados. Os dados são geralmente estruturados, organizados em tabelas relacionais e otimizados para leitura.

Por outro lado, de acordo com Armbrust et al. (2021) e Schneider et al. (2024), o Data Lake é um repositório de dados que permite armazenar grandes volumes de dados em seu formato bruto, aceitando dados estruturados, semiestruturados e não estruturados. Utiliza a abordagem schema-on-read, onde a definição do esquema ocorre apenas no momento da leitura, conferindo flexibilidade na ingestão de dados. É comumente implementado sobre sistemas de arquivos distribuídos e escaláveis, como Hadoop HDFS ou Amazon S3, utilizando formatos abertos como Apache Parquet e ORC. Apesar da flexibilidade, a ausência de controle transacional e de governança estruturada pode comprometer a qualidade e o valor analítico dos dados, levando ao risco de formação de um "data swamp".

Ademais, segundo Armbrust et al. (2021), Schneider et al. (2024) e Mazumdar et al. (2023), o Lakehouse é uma arquitetura emergente que busca integrar as vantagens do Data Warehouse — como transações ACID, controle de versões e governança de dados — com a flexibilidade e escalabilidade dos Data Lakes. Essa abordagem se apoia em formatos abertos (como Parquet e ORC) e em camadas de metadados e controle transacional, permitindo que dados brutos e processados sejam acessados e analisados em um único ambiente. O objetivo é simplificar arquiteturas analíticas, reduzir a redundância de dados e atender a cargas de trabalho variadas, desde business intelligence até machine learning.

Ainda nesse contexto, segundo Armbrust et al. (2020), Ryan et al. (2023) e Balan et al. (2022), as principais estruturas de dados utilizadas em ambientes de Big Data são baseadas em formatos colunares como Apache Parquet, ORC e Avro, armazenados sobre sistemas distribuídos como HDFS ou object stores. Essas estruturas dão suporte a dados estruturados, semiestruturados e, em menor escala, não estruturados , viabilizando operações de leitura incremental, particionamento, evolução de esquemas e versionamento. Ferramentas como Delta Lake, Apache Iceberg e Apache Hudi adotam essas estruturas para garantir performance e escalabilidade em ambientes analíticos distribuídos.

Tabela 1 - Principais tipos de tecnologias integradas em ambientes Lakehouse

Estrutura	**Descrição**
Delta Lake	Formato transacional baseado em Parquet, com suporte a ACID, schema enforcement e versionamento.
Apache Iceberg	Formato tabular open-source com controle de versão, evolução de esquema e particionamento oculto.

Apache Hudi	Estrutura para ingestão incremental e atualização de grandes volumes de dados em lakes.

Fonte: Armbrust et al. (2020), Ryan et al. (2023) e Balan et al. (2022) (Adaptado)

Para garantir compatibilidade com arquiteturas Lakehouse modernas, será considerado critério de elegibilidade que a ferramenta ofereça suporte a pelo menos uma das estruturas descritas na Tabela 1.

2.2 PERFILAMENTO DE DADOS

O perfilamento de dados (data profiling) é um processo técnico, exploratório e sistemático que tem como finalidade examinar conjuntos de dados com o intuito de extrair metadados estruturais e estatísticos. Segundo Abedjan, Golab e Naumann (2015), trata-se de uma atividade essencial nos projetos de engenharia e governança de dados, que permite obter uma visão detalhada da qualidade, da estrutura e dos padrões internos dos dados antes de qualquer uso analítico ou operacional.

Ainda segundo Abedjan, Golab e Naumann (2015), O data profiling fornece subsídios para a compreensão da estrutura real dos dados armazenados, a identificação de inconsistências e anomalias, e o reconhecimento de possíveis relações funcionais ou semânticas entre atributos. Pode ser aplicado tanto em ambientes com esquemas formalmente definidos quanto em repositórios não estruturados, sendo especialmente útil em contextos onde a documentação dos dados é inexistente, incompleta ou imprecisa.

No cenário de Big Data, o perfilamento de dados assume papel ainda mais relevante. A diversidade de fontes, a heterogeneidade dos formatos e o grande volume de informações dificultam o controle sobre a qualidade e a integridade dos dados. Além disso, o ciclo de vida dos dados em Big Data — frequentemente marcado por ingestão contínua, armazenamento distribuído e atualizações assíncronas — exige métodos que possam diagnosticar rapidamente os problemas subjacentes aos dados disponíveis.

Entretanto, conforme apontado por Abedjan, Golab e Naumann (2015), o data profiling enfrenta obstáculos específicos quando aplicado em larga escala. Um dos principais está relacionado à complexidade computacional das tarefas de análise. Operações que exigem

varreduras completas sobre colunas extensas ou que envolvem cruzamentos entre múltiplos atributos tornam-se inviáveis com o aumento exponencial do volume de dados. Para lidar com isso, técnicas como amostragem, indexação parcial e análise incremental são frequentemente necessárias, mesmo nos métodos mais rigorosos.

Outro desafio é a ausência de esquemas explícitos nos repositórios de dados modernos, como os data lakes. Nessas estruturas, os dados são armazenados com pouca ou nenhuma padronização, o que dificulta a identificação de tipos, padrões e dependências. Como destacam Abedjan, Golab e Naumann (2015), em tais contextos, o perfilamento de dados precisa assumir também o papel de descoberta estrutural (schema discovery), inferindo características que deveriam, idealmente, estar formalmente declaradas.

Segundo Abedjan, Golab e Naumann (2015), em ambientes com datasets massivos, o escopo da análise não pode ser completamente abrangente sem comprometer o desempenho do sistema. Para contornar essa limitação, é necessário adotar estratégias seletivas, como a priorização de atributos críticos ou a análise em janelas de tempo, a fim de gerar metadados úteis sem sobrecarregar o sistema de processamento.

Por fim, um ponto crítico identificado na literatura é a interpretação dos metadados gerados. Em muitos casos, os resultados do perfilamento — como cardinalidade, padrões de distribuição ou taxas de completude — não são autossuficientes. Eles devem ser contextualizados por especialistas que conheçam o domínio dos dados, pois o significado e a relevância das descobertas dependem diretamente do uso pretendido. O data profiling, nesse sentido, não substitui o conhecimento de domínio, mas o complementa com evidências empíricas sobre os dados observados.

Dessa forma, o perfilamento de dados configura-se como uma atividade central no contexto de dados em larga escala, tanto pela sua função diagnóstica quanto pelo seu papel estratégico na prevenção de problemas de qualidade e integridade. Seu uso é indispensável para qualquer aplicação analítica que se proponha a operar sobre dados massivos de forma confiável e interpretável.

2.3 TAXONOMIA DE TAREFAS DE PERFILAMENTO DE DADOS

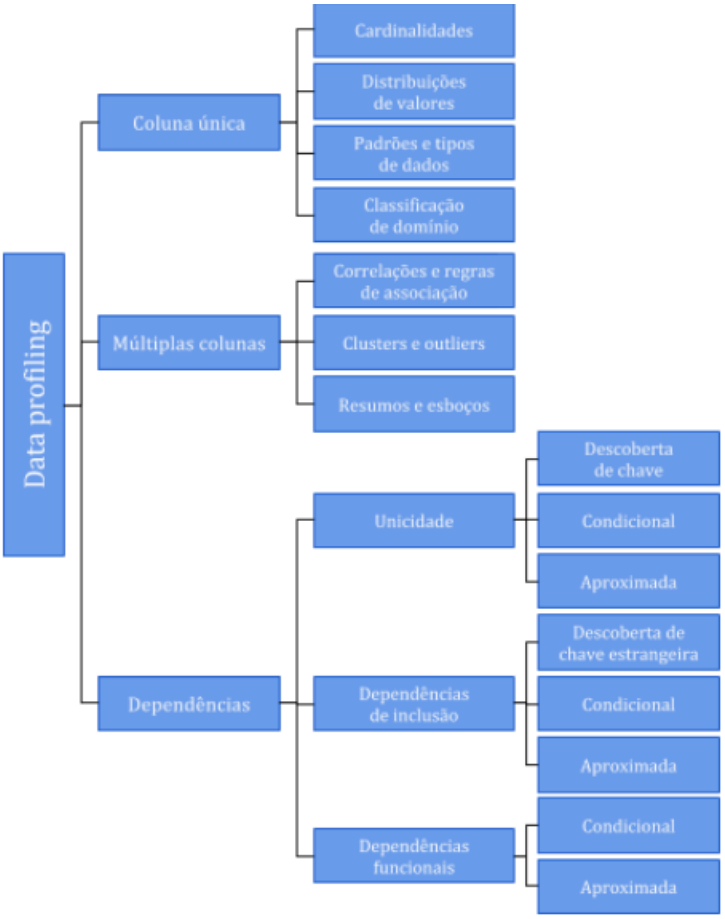
O perfilamento de dados, enquanto processo técnico, abrange um conjunto variado de tarefas que podem ser classificadas segundo seu escopo e complexidade. A estruturação sistemática dessas tarefas foi proposta por Abedjan, Golab e Naumann (2015), em uma taxonomia clássica que se tornou referência na literatura. Segundo os autores, as tarefas

podem ser agrupadas em três grandes categorias: perfilamento de coluna única, perfilamento de múltiplas colunas e descoberta de dependências.

Embora a taxonomia clássica ainda seja amplamente utilizada na literatura técnica e acadêmica, abordagens mais recentes têm proposto ampliações e reorganizações baseadas em observações práticas. Ottley et al. (2024), por exemplo, ao entrevistar analistas de dados de diferentes áreas, propuseram uma categorização mais pragmática, centrada em três eixos: caracterização descritiva, verificação de qualidade e exploração visual. Essa abordagem valoriza não apenas a estrutura dos dados, mas também o contexto da análise e o papel da interpretação humana na construção de significado.

Tais abordagens contemporâneas não invalidam a taxonomia clássica, mas a complementam, ao reconhecer que o processo de data profiling não é apenas técnico, mas também interpretativo e iterativo.

Figura 4 – Taxonomia clássica das tarefas de Data Profiling



Fonte: ABEDJAN; GOLAB; NAUMANN, 2015 (adaptado)

Essa taxonomia visa facilitar a organização do processo de análise dos dados e oferecer um caminho estruturado para extrair conhecimento confiável sobre a estrutura e o comportamento dos atributos analisados. A seguir, detalham-se os grupos que compõem essa proposta.

2.3.1 PERFILAMENTO DE COLUNA ÚNICA

Essa categoria corresponde às tarefas mais elementares do data profiling e constitui a base da maioria das análises exploratórias. Envolve o exame de colunas isoladas, com o objetivo de inferir propriedades estatísticas, estruturais e semânticas. Entre os principais elementos analisados estão:

- Cardinalidade: quantidade de valores distintos presentes na coluna. Um valor de cardinalidade muito baixo pode indicar um atributo categórico ou mesmo um campo mal preenchido; valores altos podem indicar chaves candidatas.
- Distribuições de valores: análise da frequência de ocorrência de cada valor (moda), bem como de propriedades como média, mediana e desvio padrão em colunas numéricas.
- Detecção de padrões: identificação de regularidades nos dados, como a ocorrência de datas no formato dd/mm/aaaa, e-mails, números de CPF, entre outros.
- Presença de valores nulos ou inválidos: cálculo da taxa de preenchimento, que permite estimar a completude do atributo.
- Comprimento e tipo inferido dos dados: inspeção do número médio de caracteres em atributos de texto e inferência do tipo lógico (string, inteiro, decimal, data).

De acordo com Abedjan, Golab e Naumann (2015), esse conjunto de tarefas possibilita, entre outras coisas, o diagnóstico de problemas de preenchimento, a identificação de outliers e a sugestão de padronizações de formato.

2.3.2 PERFILAMENTO DE MÚLTIPLAS COLUNAS

As tarefas desta categoria têm como foco a análise simultânea de dois ou mais atributos, com o intuito de detectar relações estatísticas ou estruturais entre eles. Esse tipo de análise é

especialmente útil em contextos em que atributos distintos possuem semântica relacionada ou complementações mútuas. Os principais aspectos analisados incluem:

- Distribuições conjuntas: comparação entre frequências de pares de atributos, como "estado" e "cidade", por meio de tabelas de contingência.
- Consistência intercolunas: avaliação de vínculos esperados entre atributos, como a consistência entre "idade" e "data de nascimento", ou entre "valor total" e a soma de "valor unitário" e "quantidade".
- Correlações numéricas: identificação de relações de dependência estatística entre colunas quantitativas, via coeficientes como Pearson ou Spearman.
- Duplicação parcial ou total: detecção de registros com valores idênticos em múltiplas colunas, o que pode indicar redundância ou falha de normalização.

Segundo Abedjan et al. (2015), o perfilamento multicolunar amplia a capacidade de compreensão dos dados ao observar o comportamento dos atributos em conjunto, e não de forma isolada.

2.3.3 DESCOBERTA DE DEPENDÊNCIAS

Essa categoria envolve tarefas mais avançadas, voltadas à identificação de regras lógicas ou estruturais que regem os dados. Tais regras podem não estar explícitas no esquema original, mas são frequentemente observadas empiricamente nas relações entre os atributos. Destacam-se:

- Dependências funcionais (FDs): relações em que o valor de um atributo determina o de outro (por exemplo, "CPF" determina "Nome"). São úteis na validação de integridade e na identificação de chaves.
- Chaves candidatas: colunas ou conjuntos de colunas que, isoladamente ou combinadas, identificam unicamente cada registro na tabela.
- Dependências de inclusão (INDs): situações em que todos os valores de uma coluna estão contidos em outra, como em relações entre tabelas mestre-detalhe.
- Dependências condicionais ou parciais: regras que se aplicam apenas a subconjuntos dos dados, como "CEP determina cidade" apenas para registros do Brasil.

- Dependências aproximadas: consideram margens de erro ou exceções, úteis quando há ruído nos dados ou inconsistências toleráveis.

Abedjan, Golab e Naumann (2015) destacam ainda que tais tarefas exigem maior poder computacional e o uso de técnicas específicas para lidar com ambiguidade e incerteza, especialmente em contextos de dados massivos e não estruturados.

3 FERRAMENTAS PARA PERFILAMENTO DE DADOS EM LARGA ESCALA

3.1 PROCESSO DE SELEÇÃO DAS FERRAMENTAS

A seleção das ferramentas de data profiling analisadas neste trabalho foi orientada por critérios metodológicos que visam garantir relevância prática, viabilidade técnica e aderência ao cenário de Big Data.

Para isso, foram utilizados os seguintes critérios de Inclusão das ferramentas:

Tabela 2 – Critérios de Seleção Utilizados na Seleção das Ferramentas

Código do Critério de Inclusão	Descrição do Critério
Critério de Seleção 1	Disponibilidade gratuita: ferramentas com versão open source ou free trial, permitindo testes com datasets reais.
Critério de Seleção 2	Reconhecimento técnico: ferramentas mencionadas em artigos especializados, revisões técnicas ou rankings de mercado;
Critério de Seleção 3	Replicabilidade experimental: facilidade de instalação e execução com dados públicos, possibilitando reprodutibilidade dos testes.
Critério de Seleção 4	Ferramenta possui suporte a pelo menos um dos seguintes tipos de dados: Delta Lake, Apache Iceberg e Apache Hudi.

Fonte: Elaborado pelo próprio autor

Para garantir uma escolha fundamentada e abrangente, realizou-se uma busca exploratória em portais especializados, sendo eles:

Tabela 3 – Plataformas de pesquisa de Ferramentas de Perfilamento de Dados

Plataforma	Descrição da Plataforma
DBMS Tools	Plataforma que cataloga e compara ferramentas de governança e análise de dados.
AltexSoft	Blog técnico com análises sobre ferramentas de data profiling.
Blog oficial do Databricks	Fonte primária para identificação de funcionalidades nativas de perfilamento de dados.

Fonte: Elaborado pelo próprio autor

Essa etapa resultou em um conjunto inicial com 13 ferramentas potencialmente relevantes, a saber: Ataccama ONE, Data Ladder, IBM InfoSphere, OpenRefine, Talend Data Fabric, Soda Core, yDataProfiling, Databricks Notebook Profiling, Dataedo, Informatica Data Profiling, DQLabs, Alation Data Catalog e Atlan.

Após a aplicação dos quatro critérios de seleção, foram selecionadas ao final cinco potenciais ferramentas que atenderam a todos os requisitos: Databricks, yDataProfiling, Dataedo.

3.2 MÉTODO DE MENSURAÇÃO FUNCIONAL DAS FERRAMENTAS

Para realizar a mensuração de ferramentas de data profiling, torna-se essencial estabelecer uma métrica sistemática de avaliação, de modo a analisar as funcionalidades oferecidas por cada solução de forma criteriosa e padronizada. Nesse sentido, o uso de um checklist funcional é uma estratégia eficaz para garantir a comparabilidade entre as ferramentas analisadas, permitindo identificar tanto suas convergências quanto suas especificidades.

Conforme apresentado por Abedjan, Golab e Naumann (2015), um dos principais desafios do perfilamento de dados está relacionado à definição clara dos resultados esperados. Em outras palavras, é necessário estabelecer quais tarefas de profiling devem ser executadas e em quais subconjuntos dos dados elas devem ser aplicadas. Muitas ferramentas requerem essa

especificação de forma explícita, enquanto outras adotam abordagens mais abrangentes, executando automaticamente diversas tarefas com o objetivo de identificar metadados relevantes.

Nesse contexto, destaca-se o trabalho de Ehrlinger e Wöß (2019), que propuseram um modelo estruturado de avaliação funcional denominado DQ Evaluation. Esse framework foi desenvolvido com base em uma revisão do estado da arte em qualidade de dados, incorporando definições e taxonomias clássicas estabelecidas por Abedjan et al. (2015, 2019). A proposta visa organizar a análise de ferramentas em três dimensões principais: (1) *data profiling*, (2) medição de qualidade de dados (*data quality measurement*) e (3) monitoramento automatizado da qualidade de dados (*automated data quality monitoring*).

A construção do modelo foi acompanhada da elaboração de um catálogo de requisitos detalhado, o qual serviu como base para um checklist funcional aplicado na prática. Utilizando critérios de exclusão predefinidos, de 267 ferramentas no total, os autores selecionaram 13 ferramentas — sendo 8 comerciais e 5 de código aberto — para análise comparativa. A aplicação do checklist permitiu mensurar o grau de suporte de cada solução às funcionalidades descritas no modelo, demonstrando sua aplicabilidade como ferramenta de avaliação padronizada.

Para os fins deste trabalho, será utilizada apenas a dimensão de data profiling do modelo DQ Evaluation, uma vez que as demais dimensões — relacionadas à medição e monitoramento da qualidade dos dados — extrapolam o escopo deste estudo. A escolha por esse framework decorre de sua fundamentação teórica consolidada, aliada à sua validação empírica por meio da aplicação em múltiplas ferramentas, o que o torna especialmente adequado para fins de comparação funcional sistemática.

Figura 5 - Catálogo de requisitos de Perfilamento de dados de ferramentas de Qualidade de Dados

Category	Sub-category	Requirement
Data Profiling (Abedjan et al., 2015, 2019)	SC – Cardinalities	(1) "Number of rows" (Abedjan et al., 2019)
		(2) Number of null values (Abedjan et al., 2019)
		(3) "Percentage of null values" (Abedjan et al., 2019)
		(4) "Number of distinct values; sometimes called 'cardinality' " (Abedjan et al., 2019)
		(5) "Number of distinct values divided by the number of rows" (Abedjan et al., 2019)
	SC - Value distributions	(6) "Frequency histograms (equi-width, equi-depth, etc.)" (Abedjan et al., 2019)
		(7) "Minimum and maximum values in a numeric column" (Abedjan et al., 2019)
		(8) "Constancy: frequency of most frequent value divided by number of rows" (Abedjan et al., 2019)
		(9) "Quartiles: 3 points that divide the (numeric) values into 4 equal groups" (Abedjan et al., 2019)
		(10) "Distribution of first digit in numeric values; to check Benford's law" (Abedjan et al., 2019)
	SC - Patterns, data types, and domains	(11) "Basic type (numeric, alphanumeric, date, time, etc.)" (Abedjan et al., 2019)
		(12) "DBMS-specific data type (varchar, timestamp, etc.)" (Abedjan et al., 2019)
		(13) Measurement of value length (minimum, maximum, average, and median) (Abedjan et al., 2019)
		(14) "Maximum number of digits in numeric values" (Abedjan et al., 2019)
		(15) "Maximum number of decimals in numeric values" (Abedjan et al., 2019)
		(16) "Histogram of value patterns (Aa9...)" (Abedjan et al., 2019)
		(17) "Generic semantic data type" (Abedjan et al., 2019) [e.g., "code, date/time, quantity, identifier" (Abedjan et al., 2019)]
		(18) "Semantic domain" (Abedjan et al., 2019) (e.g., credit card, first name, city) (Abedjan et al., 2019)
	Dependencies	(19) "Unique column combinations" (Abedjan et al., 2019) (key discovery)
		(20) "Relaxed unqiue column combinations" (Abedjan et al., 2019)
		(21) "Inclusion dependencies" (Abedjan et al., 2019) (foreign key discovery)
		(22) "Relaxed inclusion dependencies" (Abedjan et al., 2019)
		(23) "Functional dependencies" (Abedjan et al., 2019)
		(24) "Relaxed functional dependencies" (Abedjan et al., 2019)
	Advanced MC profiling	(25) Correlation analysis (Abedjan et al., 2015)
		(26) Association rule mining (Abedjan et al., 2015)
		(27) Cluster analysis (Abedjan et al., 2015)
		(28) Outlier detection (Abedjan et al., 2015)
		(29) Exact duplicate tuple detection
		(30) Relaxed duplicate tuple detection

Fonte: Ehrlinger & Wöß (2019)

Traduzindo e descrevendo cada tarefa do checklist utilizado no trabalho de Ehrlinger e Wöß (2019), teremos o seguinte checklist:

Tabela 4 - Checklist de Tarefas de Perfilamento de dados

Grupo de Tarefas	Tarefa	Descrição
Cardinalidades	Número de linhas	Total de registros presentes em uma tabela ou dataset.
	Número de valores nulos	Quantidade de valores ausentes (null) em uma coluna.
	Porcentagem de valores nulos	Proporção de valores nulos em relação ao total de registros.

	Número de valores distintos	Quantidade de valores únicos em uma coluna; também chamada de cardinalidade.
	Porcentagem de valores distintos	Proporção de valores distintos dividida pelo número total de linhas.
Distribuição de Valores	Histogramas de frequência	Representação da distribuição dos valores em uma coluna.
	Valores mínimo e máximo	Valor mínimo e máximo presentes em uma coluna numérica.
	Constância	Frequência do valor mais comum dividida pelo total de linhas.
	Quartis	Divisões dos valores numéricos em quatro grupos com quantidades iguais.
	Distribuição do primeiro dígito	Verificação da Lei de Benford no primeiro dígito de valores numéricos.
Padrões, Tipos de Dados e Domínios Semânticos	Tipos básicos	Tipo inferido de dados: numérico, alfanumérico, data, hora etc.
	Tipo de dado específico do SGBD	Tipo declarado no SGBD (ex: VARCHAR, INT).
	Comprimento do valor	Medição do comprimento dos valores (mínimo, máximo, média, mediana).
	Número de dígitos	Número máximo de dígitos em valores numéricos.
	Número de casas decimais	Número máximo de casas decimais em valores numéricos.
	Histograma de padrões de valores	Identificação de padrões estruturais nos valores (ex: Aa9...).
	Tipo de dado semântico genérico	Classificação como código, data/hora, quantidade, identificador.
	Domínio semântico	Categoria como nome próprio, cidade, cartão de crédito etc.
Dependências	Chaves candidatas (UCCs)	Descoberta de combinações únicas de colunas (possíveis chaves primárias).
	Chaves	UCCs com permissão de exceções ou duplicações ocasionais.

	candidatas relaxadas	
	Chaves estrangeiras (INDs)	Relações entre colunas de diferentes tabelas (chave estrangeira).
	Chaves estrangeiras relaxadas	Dependências de inclusão tolerando inconsistências.
	Dependências funcionais (FDs)	Relações determinísticas entre colunas.
	Dependências funcionais relaxadas	FDs que permitem algumas violações.
Perfilamento Multicoluna Avançado	Análise de correlação	Verificação de correlação estatística entre colunas.
	Regras de associação	Descoberta de padrões frequentes entre colunas.
	Análise de cluster	Agrupamento de registros por similaridade.
	Detecção de outliers	Identificação de valores significativamente discrepantes.
	Detecção de duplicatas exatas	Identificação de registros idênticos no dataset
	Detecção de duplicatas relaxadas	Identificação de registros quase idênticos com pequenas variações.

Fonte: Ehrlinger & Wöß, 2019 (Adaptado)

3.3 ANÁLISE FUNCIONAL

3.3.1 DATABRICKS

O Databricks é uma plataforma moderna de análise de dados desenvolvida sobre o Apache Spark, com o objetivo de unificar tarefas de engenharia de dados, ciência de dados, aprendizado de máquina e inteligência de negócios em um ambiente colaborativo e escalável. A ferramenta é amplamente utilizada em arquiteturas Lakehouse, oferecendo suporte nativo

ao formato Delta Lake, além de integrar-se de forma eficiente com sistemas distribuídos, como HDFS, S3 e Azure Data Lake (DATABRICKS, 2024).

A funcionalidade de data profiling foi introduzida no ambiente de notebooks a partir da versão DBR 9.1, podendo ser acessada por meio da aba gráfica “Data Profile” ou por comandos programáticos com a API *dbutils*. Essa funcionalidade permite a geração automática de perfis de colunas numéricas, textuais e temporais, incluindo estatísticas descritivas (mínimo, máximo, média, mediana, desvio padrão), frequência de valores, distribuição por histogramas e detecção de valores nulos (GAN; LEE; FORD, 2021).

A interface gráfica é um dos pontos fortes do Databricks, pois permite análise visual e interativa de grandes volumes de dados sem necessidade de escrever código, o que facilita o trabalho exploratório inicial. A identificação automática dos tipos de dados e a geração de perfis para todo o dataset — e não apenas para a amostra visível — são diferenciais positivos frente a outras soluções que operam apenas por amostragem ou sem visualização integrada.

Contudo, a plataforma apresenta limitações importantes quando analisada sob a ótica do perfilamento de dados avançado. Funcionalidades como detecção automatizada de outliers, análise de dependências funcionais, descoberta de chaves candidatas ou validação semântica entre colunas não estão disponíveis de forma nativa. Em muitos casos, é necessário recorrer a implementações personalizadas utilizando Spark SQL ou bibliotecas externas em Python, o que exige conhecimento técnico e amplia a complexidade do processo (DATABRICKS, 2024).

Além disso, o Databricks não oferece suporte direto à categorização semântica de dados (ex: identificação de nomes próprios, localidades ou códigos), nem a tarefas como análise de duplicidade aproximada ou regras de associação. Essa ausência de recursos automatizados limita seu uso em etapas posteriores do pipeline de qualidade de dados, como imputação, conformidade de esquemas ou descoberta de relacionamentos entre entidades.

Essa carência de funcionalidades mais sofisticadas pode ser explicada pela foco da plataforma na escalabilidade e integração com o ecossistema Spark, o que favorece o desempenho em ambientes distribuídos, mas limita a profundidade das análises sem programação adicional. Assim, o Databricks se mostra mais indicado para engenheiros de dados que priorizam desempenho, paralelismo e integração com pipelines analíticos, especialmente em projetos onde o perfilamento de dados serve como diagnóstico preliminar, e não como validação semântica ou estrutural completa.

3.3.2 YDATAPROFILING

O yDataProfiling, anteriormente conhecido como Pandas Profiling, é uma biblioteca de código aberto desenvolvida em Python, voltada para a geração automatizada de relatórios de perfilamento de dados. A ferramenta oferece integração direta com bibliotecas populares como Pandas e PySpark, sendo amplamente utilizada em fluxos de análise exploratória e preparação de dados em ambientes de ciência de dados (YDATA, 2025).

Segundo Clemente et al. (2023), a biblioteca permite a geração de relatórios HTML ricos em conteúdo visual, incluindo estatísticas descritivas, detecção de valores nulos, outliers, correlações, distribuição dos dados e padrões semânticos básicos. A integração com Spark DataFrames possibilita a aplicação da ferramenta em conjuntos de dados de grande volume, sem necessidade de reconfiguração do cluster Spark, o que amplia seu uso em ambientes de Big Data.

Entre seus principais diferenciais estão as funcionalidades voltadas para avaliação automática da qualidade dos dados, como alertas estatísticos, verificação de consistência, análise de distribuição e visualizações interativas. A ferramenta também suporta técnicas multivariadas como PCA e t-SNE, permitindo maior profundidade na caracterização estrutural e estatística dos dados. Além disso, o yDataProfiling fornece sugestões automáticas de tratamentos, como imputações, filtros e agrupamentos, o que torna a ferramenta bastante completa no contexto do diagnóstico de dados (ALTEXSOFT, 2025).

Entretanto, mesmo com sua cobertura funcional ampla, a ferramenta apresenta algumas limitações técnicas e operacionais. O uso intensivo de memória em conjuntos de dados muito grandes pode comprometer o desempenho, especialmente em máquinas locais ou clusters com recursos restritos. Além disso, apesar de suportar integração com Spark, algumas operações mais pesadas exigem amostragem dos dados ou podem falhar com datasets muito desbalanceados, o que requer atenção por parte do usuário (CLEMENTE ET AL., 2023).

Outra limitação observada refere-se à ausência de mecanismos robustos de validação relacional entre colunas ou tabelas, como dependências funcionais, chaves estrangeiras ou validação de integridade referencial. Tais funcionalidades são essenciais em ambientes de engenharia de dados e governança, mas ainda não são foco da ferramenta, que prioriza a visualização e o diagnóstico superficial.

Dessa forma, o yDataProfiling se destaca como uma ferramenta de alto valor para profissionais de ciência de dados e analistas exploratórios, sendo ideal para uso em etapas em que necessita de uma análise mais aprofundada do dataset, onde a rapidez na geração de

insights visuais e estatísticos é um pouco menos importante que a profundidade da análise dos datasets.

3.3.3 DATAEDO

O Dataedo é uma ferramenta comercial com versão gratuita voltada à documentação, catalogação e governança de dados, com funcionalidades integradas de perfilamento estatístico. A ferramenta possui suporte a diversos sistemas gerenciadores de banco de dados, como SQL Server, PostgreSQL, MySQL e Oracle, e é compatível com o formato Delta Lake, o que viabiliza seu uso em arquiteturas Lakehouse (DATAEDO, 2025).

Diferente das ferramentas anteriores, o foco principal do Dataedo está na documentação técnica e no mapeamento de metadados estruturais, oferecendo suporte à extração automática de esquemas, dicionários de dados, identificação de chaves primárias e estrangeiras, além da geração de diagramas ER e glossários de termos. Essas funcionalidades o tornam uma solução robusta para equipes de engenharia de dados e arquitetura que atuam com governança e catalogação.

Em relação ao perfilamento de dados, o Dataedo permite executar análises estatísticas simples, como distribuição de valores, frequência, cardinalidade, e identificação de colunas com valores nulos. A ferramenta também destaca possíveis chaves candidatas com base nos metadados existentes, facilitando a validação de integridade e estrutura do banco de dados (DBMS TOOLS, 2025).

Contudo, sua cobertura funcional para tarefas avançadas de data profiling é limitada. Funcionalidades como detecção de padrões estruturais, análise de outliers, correlações multivariadas, ou avaliação estatística da qualidade dos dados não são suportadas. Além disso, o Dataedo não oferece visualizações interativas nem alertas automatizados, o que restringe seu uso para diagnósticos visuais ou exploração interativa dos dados.

Essas características indicam que o Dataedo é mais indicado para equipes de governança de dados, compliance ou arquitetura técnica, que necessitam de controle e rastreabilidade dos ativos de dados, mas não demandam análise exploratória profunda. Sua proposta está alinhada à gestão e documentação de repositórios relacionais, e não à exploração estatística de dados massivos ou não estruturados.

Portanto, embora seja limitado em tarefas de perfilamento exploratório, o Dataedo apresenta um diferencial claro como ferramenta de documentação e inventário de dados,

sendo uma solução complementar em ecossistemas onde a rastreabilidade e a clareza estrutural são prioritárias.

3.3.4 TABELA COMPARATIVA DAS FERRAMENTAS

Com o objetivo de sintetizar os resultados da análise funcional, a Tabela 6 apresenta uma comparação entre as ferramentas Databricks, yDataProfiling e Dataedo, com base no checklist funcional adaptado de Ottley et al. (2024). A tabela contempla os principais grupos de tarefas de data profiling, indicando a presença ou ausência de suporte para cada funcionalidade em cada ferramenta.

Nota: As funcionalidades descritas foram verificadas por meio de testes práticos e complementadas com a documentação oficial da ferramenta. Capturas de tela exemplificativas estão disponíveis nos Anexos A, B e C.

Para fins de legibilidade, foram utilizadas as seguintes marcações:

Tabela 5 - Legenda Para o Checklist

Legenda	Significado
Possui	Possui a funcionalidade nativamente na ferramenta
Não Possui	Não possui a Funcionalidade na ferramenta

Fonte: Elaborado pelo próprio Autor

Abaixo, segue análise feita das três ferramentas e comparadas a partir do checklist proposto.

Tabela 6 - Checklist de Tarefas de Perfilamento de dados aplicada às três ferramentas

Grupo de Tarefas	Tarefa	Databricks	yDataProfiling	Dataedo
Cardinalidades	Número de linhas	Possui	Possui	Possui
	Número de valores nulos	Possui	Possui	Possui
	Porcentagem de valores nulos	Possui	Possui	Possui
	Número de valores distintos	Possui	Possui	Possui

	Porcentagem de valores distintos	Possui	Possui	Possui
Distribuição de Valores	Histogramas de frequência	Possui	Possui	Possui
	Valores mínimo e máximo	Possui	Possui	Possui
	Constância	Possui	Possui	Possui
	Quartis	Possui	Possui	Não Possui
	Distribuição do primeiro dígito	Não Possui	Possui	Não Possui
Padrões, Tipos de Dados e Domínios Semânticos	Tipos básicos	Possui	Possui	Possui
	Tipo de dado específico do SGBD	Possui	Não Possui	Possui
	Comprimento do valor	Não Possui	Possui	Possui
	Número de dígitos	Possui	Possui	Possui
	Número de casas decimais	Possui	Possui	Possui
	Histograma de padrões de valores	Possui	Possui	Possui
	Tipo de dado semântico genérico	Possui	Possui	Possui
	Domínio semântico	Não Possui	Não Possui	Não Possui
Dependências	Chaves candidatas (UCCs)	Possui	Possui	Possui
	Chaves candidatas relaxadas	Não Possui	Não Possui	Não Possui
	Chaves estrangeiras (INDs)	Possui	Não Possui	Possui
	Chaves estrangeiras relaxadas	Não Possui	Não Possui	Não Possui
	Dependências funcionais (FDs)	Não Possui	Possui	Não Possui
	Dependências funcionais relaxadas	Não Possui	Não Possui	Não Possui
Perfilamento Multicoluna Avançado	Análise de correlação	Não Possui	Possui	Não Possui
	Regras de associação	Não Possui	Não Possui	Não Possui
	Análise de cluster	Não Possui	Não Possui	Não Possui

	Detecção de outliers	Não Possui	Não Possui	Não Possui
	Detecção de duplicatas exatas	Não Possui	Possui	Não Possui
	Detecção de duplicatas relaxadas Avançado	Não Possui	Não Possui	Não Possui

Com base na análise funcional das ferramentas yDataProfiling, Databricks e Dataedo, observou-se que a cobertura de funcionalidades de perfilamento de dados varia significativamente entre elas. Essa variação reflete não apenas os propósitos originais de cada ferramenta, mas também seu público-alvo e o nível de profundidade desejado nas tarefas de diagnóstico, exploração e avaliação da qualidade dos dados.

A ferramenta yDataProfiling apresentou a maior cobertura funcional no checklist proposto, destacando-se nas tarefas de caracterização estatística, análise de distribuição, detecção de padrões e visualização multivariada com técnicas como PCA e t-SNE. Além disso, conforme Clemente et al. (2023), a ferramenta incorpora alertas estatísticos e sugestões automatizadas, o que a torna particularmente adequada para atividades de preparação de dados, ciência de dados e análise preditiva.

O Databricks, por sua vez, apresentou uma cobertura mais restrita em relação à funcionalidades mais avançadas. No entanto, de acordo com Gan, Lee e Ford (2021), destaca-se pela eficiência no processamento de grandes volumes de dados, oferecendo funcionalidades básicas de perfilamento, como estatísticas descritivas, histogramas e contagem de valores nulos. Conforme documentação da própria plataforma (DATABRICKS, 2024), tarefas mais complexas, como a descoberta de chaves candidatas ou validações semânticas, exigem a utilização de Spark SQL ou implementações customizadas. Dessa forma, a ferramenta é mais indicada para engenheiros de dados que atuam em ambientes distribuídos e necessitam de agilidade na análise exploratória.

A ferramenta Dataedo, por sua vez, diferencia-se por seu foco na documentação técnica e catalogação de dados, oferecendo suporte consistente à análise de metadados, identificação de chaves primárias e estrangeiras e geração de diagramas ER. De acordo com Dataedo (2025), a ferramenta possui cobertura limitada em tarefas avançadas de perfilamento, como análise de padrões, correlações multivariadas e avaliação da qualidade estatística dos dados. Apesar disso, mostra-se uma solução sólida para equipes de governança de dados e

arquitetura, que priorizam rastreabilidade e organização estrutural sobre exploração estatística.

Dessa forma, conclui-se que o yDataProfiling é a ferramenta mais completa entre as avaliadas, sendo recomendada para contextos que exigem detalhamento estatístico, detecção de anomalias e preparação de dados para modelagem analítica. O Databricks se destaca pela escalabilidade e integração com o ecossistema Spark, sendo mais adequado para exploração inicial de dados em larga escala. Já o Dataedo atende bem às necessidades de documentação e controle de metadados, embora seja menos apropriado para análises estatísticas aprofundadas ou diagnósticos de qualidade em ambientes analíticos.

3.4 CRITÉRIOS QUALITATIVOS DE ADOÇÃO E INTEGRAÇÃO

Para além da cobertura funcional estrita das tarefas de perfilamento de dados, a adoção prática de uma ferramenta em ambientes corporativos e acadêmicos é influenciada por fatores qualitativos que impactam diretamente sua viabilidade operacional. Entre esses fatores, destacam-se a facilidade de instalação, a curva de aprendizado, a integração com pipelines de dados e a qualidade da documentação disponível.

A facilidade de instalação refere-se à simplicidade do processo de configuração inicial da ferramenta, incluindo requisitos de infraestrutura, dependências e compatibilidade com diferentes sistemas operacionais. Uma instalação simplificada reduz barreiras de entrada e agiliza o início das atividades de perfilamento.

A curva de aprendizado diz respeito ao tempo e esforço necessários para que usuários, com diferentes níveis de experiência, consigam explorar plenamente as funcionalidades da ferramenta. Uma curva reduzida favorece adoção mais rápida e maior produtividade, especialmente em equipes heterogêneas.

A integração com pipelines de dados avalia a capacidade da ferramenta de se conectar e operar de forma contínua em fluxos de processamento já estabelecidos, como notebooks interativos, orquestradores de workflows (Apache Airflow, dbt) ou pipelines distribuídos em Apache Spark. Essa característica é particularmente relevante em projetos que requerem automação e reprodutibilidade.

Por fim, a documentação compreende a disponibilidade, organização e abrangência de materiais de suporte oficial, como manuais, tutoriais, guias de boas práticas e referências de API. Documentação completa e atualizada é fundamental para reduzir dependência de suporte técnico e facilitar a resolução de problemas.

A Tabela 7 resume essas características qualitativas das ferramentas analisadas com base na experiência prática durante os testes, documentação oficial disponível e aspectos técnicos observados no processo de configuração e execução

Tabela 7 - Critérios qualitativos de adoção e integração das ferramentas analisadas

Ferramenta	Facilidade de Instalação	Curva de Aprendizado	Integração com Pipelines	Documentação
Databricks	Alta	Média	Alta	Completa
yDataProfiling	Alta	Baixa	Alta	Ampla
Dataedo	Média	Média	Baixa	Clara

Fonte: Elaborado pelo autor (2025).

No caso do Databricks, a instalação é simplificada devido ao modelo baseado em nuvem, porém a utilização exige familiaridade prévia com o ecossistema Apache Spark, o que eleva moderadamente a curva de aprendizado. Sua integração com pipelines é considerada alta, com suporte nativo a notebooks, APIs e formatos modernos como Delta Lake, além de documentação oficial extensa e detalhada.

O yDataProfiling apresenta instalação rápida via gerenciadores de pacotes Python, exigindo conhecimentos básicos em bibliotecas como Pandas ou PySpark. Sua curva de aprendizado é baixa, permitindo uso imediato, e a integração com pipelines é favorecida pelo suporte a ambientes interativos e execução distribuída. A documentação é ampla e inclui exemplos práticos, guias de referência e API detalhada.

Por sua vez, o Dataedo requer instalação local, com maior número de etapas e dependências, resultando em avaliação média para a facilidade de instalação. A curva de aprendizado também é média, pois, embora a interface seja intuitiva, a exploração plena de suas funções de governança demanda tempo. A integração com pipelines é baixa devido à ausência de SDKs ou APIs para automação, restringindo seu uso a operações manuais. A documentação é clara e bem estruturada, abrangendo desde a instalação até guias de uso avançado.

Esses critérios não substituem a análise funcional apresentada anteriormente, mas a complementam ao fornecer uma perspectiva mais realista sobre a adoção das ferramentas em diferentes contextos. Tal abordagem é coerente com práticas recomendadas em engenharia de dados, que consideram não apenas a capacidade técnica da solução, mas também seu custo de

implementação e facilidade de inserção em ecossistemas existentes (EHRLINGER; WÖß, 2019).

3.5 FERRAMENTAS EMERGENTES NÃO AVALIADAS

Além das três ferramentas analisadas neste estudo, foram identificadas outras soluções de perfilamento e qualidade de dados amplamente reconhecidas, como Talend Data Fabric, Informatica Data Quality, Great Expectations, Amazon Deequ, OpenMetadata, Soda e Monte Carlo. Essas ferramentas oferecem recursos avançados, incluindo definição programática de regras de qualidade (expectations as code), integração com pipelines de orquestração e ambientes de integração e entrega contínua (CI/CD), geração automatizada de relatórios e dashboards, e monitoramento contínuo da integridade e completude dos dados.

Apesar de sua relevância, não foram incluídas nos testes práticos por apresentarem uma ou mais das seguintes restrições: ausência de versão gratuita plenamente funcional, suporte limitado a formatos compatíveis com arquiteturas Lakehouse (como Delta Lake), dependência de infraestrutura proprietária (por exemplo, AWS para o Amazon Deequ) ou complexidade de instalação incompatível com o ambiente padronizado adotado neste trabalho.

O não aprofundamento dessas ferramentas reflete critérios metodológicos voltados à reprodutibilidade e comparabilidade dos resultados, e não sua falta de importância. Recomenda-se que pesquisas futuras incluam a avaliação dessas soluções, especialmente em contextos que demandem integração com dados reais, análise de ambientes multi-tabela e implementação de métricas contínuas de observabilidade.

4 ANÁLISE DE DESEMPENHO DE FERRAMENTAS PARA PERFILAMENTO DE DADOS EM LARGA ESCALA

No contexto do Big Data, a característica de volume desempenha um papel central tanto na definição quanto nos desafios operacionais de tratamento e análise dos dados.

Ao focar especificamente no V de volume, observa-se que esse fator tem implicações significativas no desempenho dos algoritmos de data profiling. Segundo Abedjan, Golab e Naumann (2015), a complexidade computacional dos métodos de perfilamento está diretamente ligada ao número de linhas e colunas do conjunto de dados. A ordenação dos dados, por exemplo, é uma operação recorrente, enquanto muitas tarefas exigem a análise combinatória entre colunas — o que resulta em uma escalabilidade exponencial em relação ao número de atributos.

Ainda de acordo com os autores, esse cenário é agravado quando se trata de Big Data, cujas propriedades — volume elevado, alta velocidade de geração e diversidade estrutural — dificultam a aplicação de técnicas tradicionais de análise. A necessidade de armazenar, consultar e integrar grandes volumes de dados gera custos computacionais consideráveis, o que reforça a importância de aplicar técnicas eficazes e escaláveis de data profiling. Neste sentido, torna-se indispensável considerar não apenas a precisão, mas também a eficiência e capacidade de escalabilidade das ferramentas empregadas, especialmente em ambientes distribuídos ou baseados em disco, como é o caso de arquiteturas baseadas em Lakehouse ou Delta Lake.

Considerando essa base teórica, foram selecionados dois conjuntos de dados com tamanhos distintos — um de menor volume e outro com grande escala — com o objetivo de avaliar comparativamente o desempenho das ferramentas de perfilamento forma individual frente a diferentes níveis de complexidade computacional. Tal escolha se justifica pelos apontamentos de Abedjan, Golab e Naumann (2015), que destacam a escalabilidade como um dos principais desafios no contexto de data profiling, especialmente em tarefas que exigem varredura de colunas e ordenações, cujos custos crescem exponencialmente com o aumento do volume de dados.

A escolha desses datasets considerou os fundamentos teóricos dos modelos relacionais, conforme descritos por Elmasri e Navathe (2010) e por Silberschatz, Korth e Sudarshan (2011). De acordo com esses autores, uma tabela relacional é composta por um conjunto de tuplas (linhas) e atributos (colunas), sendo que o volume lógico de dados é definido pelo produto entre essas duas dimensões: $N \times M$, em que N representa o número de

tuplas e M o número de atributos. Esse modelo estabelece uma base conceitual para mensurar o volume de dados armazenados em uma tabela e, consequentemente, compreender o impacto da escalabilidade nas tarefas de data profiling.

4.1 ESCOLHA DOS CONJUNTOS DE DADOS

Para realizar os testes de desempenho das ferramentas de data profiling, foram selecionados dois conjuntos de dados públicos amplamente utilizados na literatura científica. A escolha foi guiada pela necessidade de representar diferentes escalas de volume de dados, conforme discutido na fundamentação teórica anterior.

O primeiro conjunto de dados utilizado foi o “CSE-CIC-IDS2018”, desenvolvido pelo Canadian Institute for Cybersecurity. Esse dataset é composto por 2.097.145 linhas e 80 colunas, derivadas de dois arquivos CSV principais (02-14-2018.csv e 02-15-2018.csv). Trata-se de um dataset de detecção de intrusão em redes computacionais, frequentemente utilizado para testar algoritmos de aprendizado de máquina supervisionado em ambientes distribuídos com uso de Apache Spark. Em estudo recente, esse dataset foi empregado para avaliar o desempenho de algoritmos como Logistic Regression, SVM, Decision Trees e Naive Bayes em clusters Spark (BRAHMANE; KRISHNA, 2021). Sua alta dimensionalidade e grande volume o tornam adequado para avaliar a escalabilidade das ferramentas analisadas neste trabalho.

O segundo conjunto de dados adotado foi o CoverType, oriundo da UCI Machine Learning Repository. Composto por 581.012 linhas e 54 atributos, esse dataset contém registros de sensoriamento ambiental coletados no Roosevelt National Forest, Colorado (EUA). Seu objetivo é prever o tipo de cobertura florestal entre sete categorias, com base em atributos como elevação, declividade, tipo de solo e incidência de luz solar. Foi utilizado por Brahmane e Krishna (2021) em estudos de classificação com técnicas de deep learning e arquiteturas distribuídas baseadas em Spark. Apesar de menor em volume em comparação ao CSE-CIC-IDS2018, o CoverType apresenta complexidade estrutural suficiente para servir como contraponto na comparação entre desempenho das ferramentas de profiling.

O resumo das características de cada dataset está representado abaixo:

Tabela 8 – Características dos datasets utilizados para teste de desempenho

Dataset	Quantidade de Linhas	Quantidade de Colunas	Relação entre Linhas	Relação entre Colunas	Relação Final (Linhas × Colunas)
CSE-CIC-IDS2018	2.097.145	80	3,60×	1,48×	5,33×
CoverType	581.012	54	1×	1×	1×

Fonte: Elaborado pelo autor (2025)

Para efeito de referência, o dataset CoverType foi considerado como base (1×) e, a partir disso, calculou-se a proporção relativa do conjunto CSE-CIC-IDS2018 em termos de número de linhas, colunas e da combinação entre ambas (linhas × colunas). Essa análise evidencia a diferença significativa de escala entre os dois conjuntos de dados, o que justifica sua escolha para avaliar o impacto do volume de dados no desempenho das ferramentas de data profiling. O CSE-CIC-IDS2018 apresenta um volume potencial de dados mais de cinco vezes maior que o CoverType, tornando-o adequado para testar a escalabilidade computacional das soluções avaliadas.

Cabe destacar que o suporte a formatos estruturais modernos, como Delta Lake, Apache Iceberg e Apache Hudi, foi adotado como critério de elegibilidade por representar um indicativo de maturidade tecnológica das ferramentas frente aos desafios impostos pelo ecossistema Big Data. No entanto, os testes práticos de desempenho foram conduzidos com arquivos nos formato CSV, amplamente utilizados em ambientes analíticos. O uso de CSV foi uma limitação técnica, e que a avaliação de compatibilidade com formatos Lakehouse não foi empírica, apenas declarativa.

4.2 EXECUÇÃO DAS FERRAMENTAS

Segundo Abedjan et al. (2015), a complexidade computacional dos algoritmos de data profiling depende diretamente do volume de dados e da quantidade de colunas analisadas, sendo que muitas tarefas exigem varredura completa das combinações possíveis entre atributos. Essas operações, como ordenações e verificações de dependências funcionais, podem ter custo exponencial em relação ao número de colunas, o que afeta diretamente o tempo de execução das ferramentas. Ainda de acordo com os autores, o tempo necessário para

a execução das tarefas pode variar de minutos a horas, dependendo do volume dos dados e da profundidade da análise.

Com base nessa abordagem, este trabalho optou por adotar o tempo de execução como principal métrica para avaliação do desempenho entre as ferramentas analisadas. Embora outros recursos, como uso de CPU ou memória, também sejam relevantes, o tempo total de execução oferece uma medida comparativa objetiva e reprodutível de forma inicial, especialmente em contextos onde se deseja observar o impacto do volume de dados sobre a escalabilidade dos sistemas.

Todos essas execuções foram executadas em um computador com as seguintes configurações:

- Intel Core I5 da décima segunda geração.
- 32GB RAM
- SSD 1TB NVME 2.0

E em relação às plataformas/softwares vinculados à execução:

- Microsoft SQL Studio 2019 (Banco de Dados vinculado ao software do Dataedo)
- Plataforma própria do Databricks (Databricks Profiling - Free Trial)
- JupyterLab Anaconda Navigator (yDataProfiling)

Não foram aplicadas otimizações específicas como cache ou tuning de execução para manter a comparabilidade experimental e refletir o comportamento padrão das ferramentas.

A Tabela 9 apresenta o tempo de execução (em segundos) das três ferramentas avaliadas — Databricks, yDataProfiling e Dataedo — em dois datasets com diferentes volumes de dados: CSE-CIC-IDS2018 (maior volume) e CoverType (menor volume).

Tabela 9 – Resultados do teste de desempenho

Tempo de Execução em cada ferramenta (em segundos)			
Dataset	**Databricks**	**yDataProfiling**	**Dataedo**
CSE-CIC-IDS2018	210.00	837.49	3055.63
Covertype	59.18	119.69	76.47

Fonte: Elaborado pelo autor (2025)

Observa-se que, conforme o volume de dados aumenta, o tempo de execução das ferramentas cresce de forma significativa, evidenciando o impacto do fator volume — um dos principais "Vs" do Big Data — sobre a escalabilidade dos algoritmos de data profiling. Abaixo, apresenta-se a variação percentual entre os tempos de execução nos dois datasets:

- Databricks: apresentou um aumento de 254,84% no tempo de execução ao processar o dataset maior em comparação ao menor.
- yDataProfiling: registrou um acréscimo de 599,71% entre os dois cenários.
- Dataedo: teve o maior impacto, com um aumento de 3.895,80%, indicando significativa limitação frente a grandes volumes de dados.

Estes resultados corroboram a literatura, que destaca a complexidade computacional das tarefas de data profiling em função do número de tuplas (linhas) e atributos (colunas), especialmente quando são necessárias operações de ordenação, varredura de colunas e descoberta de padrões (ZHANG et al., 2015; ABEDJAN et al., 2015). A ferramenta Databricks demonstrou maior estabilidade e desempenho escalável, sendo a mais eficiente no cenário com grande volume de dados, enquanto o datado apresentou sensível degradação de performance, indicando menor capacidade de lidar com ambientes distribuídos e arquivos em larga escala.

4.3 LIMITAÇÕES DO CONJUNTO DE TESTES PRÁTICOS

A análise de desempenho e funcionalidade apresentada neste trabalho apresenta limitações metodológicas que precisam ser consideradas na interpretação dos resultados.

A primeira limitação refere-se ao escopo funcional distinto entre as ferramentas analisadas. Cada solução executa, por padrão, um conjunto próprio de tarefas de perfilamento, o que inviabiliza a comparação estritamente equitativa dos tempos absolutos observados. O yDataProfiling, por exemplo, realiza análises adicionais, como detecção de outliers e aplicação de análise de componentes principais, enquanto o Databricks se restringe a estatísticas descritivas básicas e o Dataedo concentra-se na extração de metadados estruturais. Essa disparidade metodológica influencia diretamente os tempos de execução, de modo que os resultados devem ser compreendidos como indicativos do comportamento observado nas condições testadas, e não como uma métrica definitiva de eficiência em condições

controladas. Portanto, caracteriza-se essa análise de forma informativa, não comparativa. Cabe considerar os melhores cenários para cada situação de acordo com o resultado.

Outra limitação está relacionada ao número reduzido de conjuntos de dados utilizados nos experimentos. Foram empregados apenas dois datasets públicos, CoverType e CSE-CIC-IDS2018, que, embora possuam diferenças relevantes em volume e estrutura, não representam a diversidade de formatos, tamanhos e complexidades encontrada em cenários reais. Essa restrição reduz a possibilidade de generalização dos resultados para bases de dados com características distintas, como registros predominantemente textuais, dados semiestruturados — por exemplo, em formato JSON ou XML — ou dados multimodais.

Além disso, optou-se pela utilização exclusiva do formato CSV nos testes, devido à sua simplicidade e ampla compatibilidade. Contudo, considerando que todas as ferramentas avaliadas declaram suporte a formatos colunares modernos, como Parquet e ORC, bem como integração com arquiteturas Lakehouse, a ausência de experimentos práticos nesses formatos impossibilita a avaliação empírica de sua performance e compatibilidade em contextos distribuídos avançados.

Por fim, os testes realizados limitaram-se a bases tabulares isoladas, não contemplando cenários relacionais compostos por múltiplas tabelas interligadas por chaves primárias e estrangeiras, tampouco dependências funcionais complexas entre entidades. Tal ausência pode ter levado à subestimação do potencial de ferramentas que apresentam recursos específicos para perfilamento relacional, como é o caso do Dataedo.

Diante dessas considerações, conclui-se que os resultados obtidos refletem as condições específicas dos experimentos conduzidos e não devem ser extrapolados de forma irrestrita para outros contextos, formatos ou volumes de dados.

5 CONSIDERAÇÕES FINAIS

O presente trabalho teve como objetivo realizar uma análise funcional e de desempenho de três ferramentas de data profiling aplicadas a ambientes de Big Data, com ênfase em arquiteturas compatíveis com o modelo Lakehouse. A partir da aplicação de um checklist funcional fundamentado na literatura, foi possível comparar as funcionalidades oferecidas por cada solução, identificando seus pontos fortes, limitações e contextos de uso mais adequados.

Os resultados obtidos indicam que as ferramentas avaliadas apresentam características que as tornam mais adequadas a diferentes perfis profissionais. O Databricks se mostrou mais indicado para engenheiros de dados que atuam em cenários de alto volume e demandam integração com pipelines Spark; o yDataProfiling demonstrou-se mais apropriado para cientistas de dados que necessitam de recursos abrangentes para análise exploratória, detecção de padrões e geração de insights; e o Dataedo evidenciou maior utilidade para analistas de governança e metadados, sendo eficaz na documentação, elaboração de dicionários de dados e controle de integridade referencial. A compreensão dessas distinções pode contribuir para uma adoção mais estratégica das ferramentas, otimizando custos, tempo de resposta e alinhamento técnico entre os membros da equipe de dados.

Constatou-se que o data profiling é uma etapa fundamental na engenharia de dados, atuando como suporte à tomada de decisão ao fornecer metadados sobre a estrutura, a qualidade e a distribuição dos dados. Esse processo contribui para diagnósticos mais precisos, integração eficiente e prevenção de erros analíticos.

Uma limitação importante está relacionada à interpretação dos resultados gerados, pois, como apontado por Abedjan, Golab e Naumann (2015), os metadados extraídos não asseguram, por si só, a identificação correta de chaves, domínios ou relações. A análise adequada exige conhecimento técnico e domínio do contexto dos dados, reforçando a importância de especialistas na interpretação. Ademais, cada ferramenta adota estratégias distintas para lidar com volume e desempenho, como amostragem, execução paralela e reaproveitamento de resultados, o que, embora útil, pode introduzir imprecisões em ambientes complexos.

Como contribuição metodológica, ressalta-se que o checklist funcional empregado é replicável e pode ser adaptado para avaliar outras ferramentas. Considera-se, portanto, que este estudo oferece uma base sólida para investigações acadêmicas e para aplicações práticas no campo da engenharia e governança de dados.

6 TRABALHOS FUTUROS

O presente estudo pode ser ampliado em diversas direções, de modo a aprofundar a compreensão sobre o perfilamento de dados em ambientes de larga escala. Uma possibilidade consiste em estender a análise para novas soluções, tanto de código aberto quanto comerciais, incluindo ferramentas voltadas à observabilidade de dados em tempo real e à integração com pipelines de qualidade de dados contínua. Essa ampliação permitiria uma visão mais abrangente do ecossistema tecnológico, contemplando funcionalidades emergentes que não foram avaliadas nesta pesquisa.

Outra oportunidade relevante diz respeito à avaliação das ferramentas em formatos de dados distintos do CSV, como Parquet e ORC, lidos a partir de camadas Delta Lake ou estruturas equivalentes. Essa abordagem possibilitaria verificar, de forma prática, a aderência declarada a arquiteturas Lakehouse e observar eventuais variações de desempenho e compatibilidade em contextos distribuídos.

Também se recomenda a inclusão de cenários relacionais mais complexos, compostos por múltiplas tabelas interligadas, com vistas a examinar a capacidade das ferramentas em detectar chaves primárias e estrangeiras, avaliar a consistência entre tabelas e identificar dependências funcionais.

Além disso, sugere-se que trabalhos futuros incorporem métricas complementares, como consumo de recursos, impacto no desempenho de pipelines e precisão das análises realizadas. A adoção de abordagens qualitativas, por meio de entrevistas ou questionários com usuários experientes, também poderia fornecer percepções adicionais sobre a facilidade de uso, a curva de aprendizado e a adequação das ferramentas a diferentes contextos profissionais.

Adicionalmente, seria pertinente avaliar ferramentas que não puderam ser instaladas ou testadas neste estudo por meio da análise de documentação técnica ou de relatos de uso, de forma a ampliar a base comparativa e incluir soluções líderes de mercado.

Embora este trabalho não tenha adotado um escopo de execução controlado, permanece como possibilidade para pesquisas posteriores a realização de experimentos padronizados, restringindo todas as ferramentas a um conjunto comum de tarefas de perfilamento, como contagem de linhas, detecção de valores nulos, cálculo de cardinalidade e identificação de tipos básicos. Tal abordagem permitiria comparações mais justas entre as soluções e contribuiria para a análise de sua eficiência em condições equivalentes.

REFERÊNCIAS

ABEDJAN, Ziawasch; GOLAB, Lukasz; NAUMANN, Felix. Profiling relational data: a survey. The VLDB Journal, Heidelberg: Springer, v. 24, n. 4, p. 557–581, 2015. DOI: https://doi.org/10.1007/s00778-015-0389-6.

ABEDJAN, Z.; GOLAB, L.; NAUMANN, F. Profiling relational data: A survey. The VLDB Journal, v. 24, n. 4, p. 557–581, 2015.

ABEDJAN, Z. et al. Detecting data anomalies. In: Proceedings of the 2019 International Conference on Management of Data (SIGMOD). 2019. p. 1007–1012.

ALTEXSOFT. Data profiling tools to build trust & improve AI-readiness. 2025. Disponível em: https://www.altexsoft.com/blog/data-profiling. Acesso em: 24 jul. 2025.

ARMBRUST, Michael et al. Delta Lake: high-performance ACID table storage over cloud object stores. In: Proceedings of the VLDB Endowment, v. 13, n. 12, p. 3411–3424, 2020.

ARMBRUST, Michael et al. Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics. In: CIDR 2021 – Conference on Innovative Data Systems Research. 2021. Disponível em: https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf. Acesso em: 23 jul. 2025.

BRAHMANE, Anilkumar V.; KRISHNA, B. Chaitanya. Big data classification using deep learning and Apache Spark architecture. Neural Computing and Applications, Springer, 2021. DOI: https://doi.org/10.1007/s00521-021-06145-w.

CLEMENTE, Fabiana et al. ydata-profiling: Accelerating data-centric AI with high-quality data. Seattle, WA: YData Labs Inc., 2023.

DATABRICKS. Notebooks: display and visualization. Databricks Documentation, 2024. Disponível em: https://docs.databricks.com/notebooks/notebooks-use.html. Acesso em: 16 jul. 2025.

DATAEDO. Data profiling in Desktop. Disponível em: https://docs.dataedo.com/data-governance/data-profiling/data-profiling-in-desktop. Acesso em: 16 jul. 2025.

DBMS TOOLS. Data profiling tools that support Delta Lake. Disponível em: https://dbmstools.com/categories/data-profiling-tools/delta-lake. Acesso em: 24 jul. 2025.

EHRLINGER, L.; WÖß, W. A survey of data quality measurement and monitoring tools. arXiv preprint, arXiv:1907.08138, 2019. Disponível em: https://arxiv.org/abs/1907.08138. Acesso em: 23 jul. 2025.

ELMASRI, Ramez; NAVATHE, Shamkant B. Fundamentos de sistemas de bancos de dados. 6. ed. São Paulo: Pearson Addison Wesley, 2010.

GAN, Edward; LEE, Moonsoo; FORD, Austin. Simplify data exploration with data profiling in Databricks Notebooks. Databricks, 7 dez. 2021. Disponível em:

https://www.databricks.com/blog/2021/12/07/simplify-data-exploration-with-data-profiling-in-databricks-notebooks.html. Acesso em: 1 jun. 2025.

GAN, Edward; LEE, Moonsoo; FORD, Austin. Introducing data profiles in the Databricks notebook. Databricks, 2021. Disponível em: https://www.databricks.com/blog/2021/12/07/introducing-data-profiles-in-the-databricks-notebook.html. Acesso em: 24 jul. 2025.

KUMAR, Abhishek; VERMA, Deepti. Performance evaluation of Apache Spark MLlib algorithms on an intrusion detection dataset. arXiv preprint, 2022. Disponível em: https://arxiv.org/abs/2212.05269. Acesso em: 16 jul. 2025.

LOSHIN, David. Chapter 14 – Data profiling. In: LOSHIN, David. Data quality: the accuracy dimension. 1. ed. Burlington: Morgan Kaufmann; Elsevier, 2011. cap. 14, p. 287–324.

MAZUMDAR, Dipankar; HUGHES, Jason; ONOFRÉ, JB. The Data Lakehouse: Data Warehousing and More. 2023. Disponível em: https://arxiv.org/abs/2310.08697. Acesso em: 23 jul. 2025.

OTTLEY, Alvitta; MORGAN, Josh; CHEN, Yifan; STASKO, John. Tasks and visualizations used for data profiling: a survey and interview study. IEEE Transactions on Visualization and Computer Graphics, New York: IEEE, 2024. DOI: 10.1109/TVCG.2024.1234567.

SCHNEIDER, Jan et al. Assessing the Lakehouse: Analysis, Requirements and Definition. In: Proceedings of the 25th International Conference on Enterprise Information Systems (ICEIS 2023). p. 44–56. DOI: 10.5220/0011840500003467. Acesso em: 23 jul. 2025.

SCHNEIDER, Jan et al. The Lakehouse: State of the Art on Concepts and Technologies. SN Computer Science, v. 5, n. 449, 2024. DOI: https://doi.org/10.1007/s42979-024-02737-0. Acesso em: 23 jul. 2025.

SHARDEN, Larissa T. Chapter 10 – Data profiling. In: TURBAN, Efraim; SHARDEN, Larissa T.; DILLON, Justin; KING, David. Business intelligence: a managerial approach. Amsterdam: Elsevier, 2013. cap. 10, p. 225–248.

SILBERSCHATZ, Abraham; KORTH, Henry F.; SUDARSHAN, S. Database system concepts. 6. ed. New York: McGraw-Hill, 2011.

SOLARMAINFRAME. Intrusion Detection Dataset - CSE-CIC-IDS2018. Kaggle, 2025. Disponível em: https://www.kaggle.com/datasets/solarmainframe/ids-intrusion-csv. Acesso em: 16 jul. 2025.

TALEB, Ikbal; SERHANI, Mohamed Adel; DSSOULI, Rachida. Big data quality: a data quality profiling model. In: WORLD CONGRESS ON SERVICES, 2019, Cham. Anais [...]. Cham: Springer International Publishing, 2019. p. 61–77.

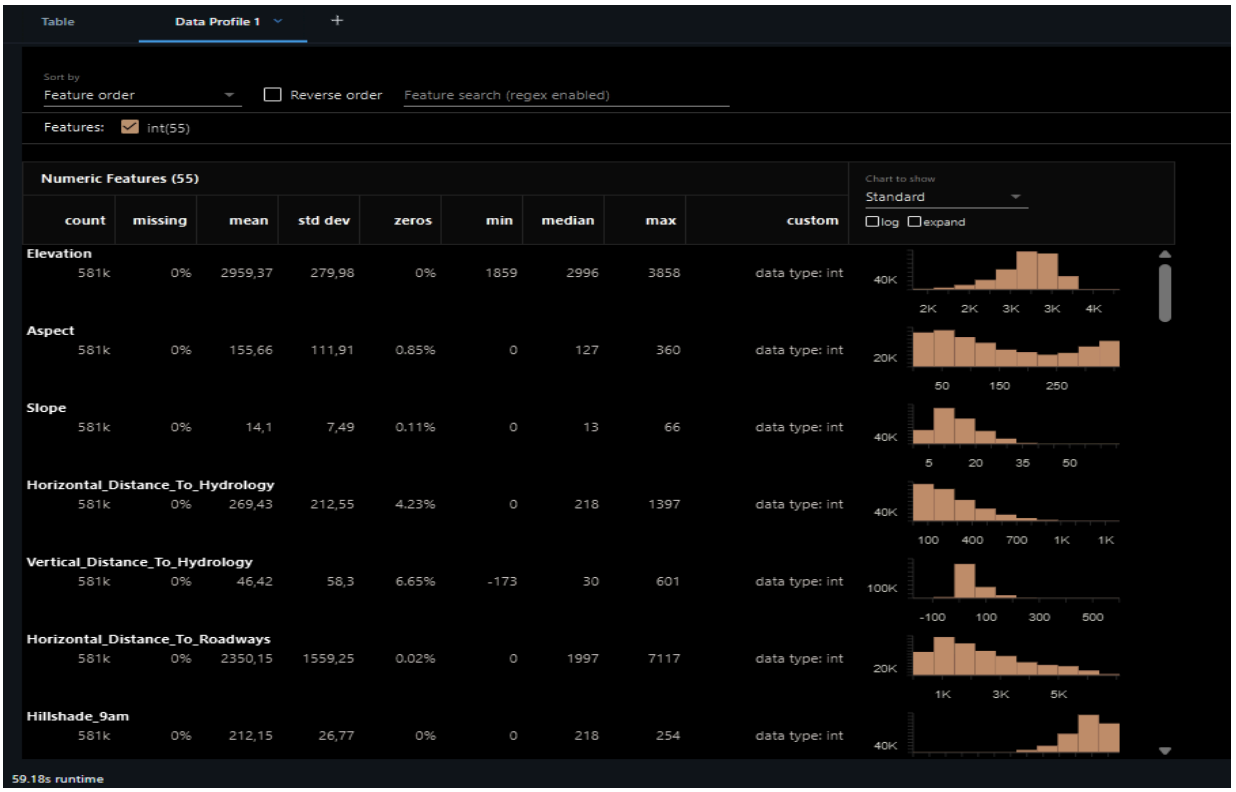
UCI MACHINE LEARNING REPOSITORY. Forest CoverType Dataset. 2025. Disponível em: https://archive.ics.uci.edu/dataset/31/covertype. Acesso em: 16 jul. 2025.

YDATA. Data profiling – Concepts. 2025. Disponível em: https://docs.profiling.ydata.ai/latest/getting-started/concepts/#preview-data. Acesso em: 16 jul. 2025.

ZHANG, Yudong; LUO, Yao. Data profiling technology of data governance regarding big data: review and rethinking. Procedia Computer Science, Amsterdam: Elsevier, v. 91, p. 659–668, 2016. DOI: 10.1016/j.procs.2016.07.166.

ANEXO A – EVIDÊNCIAS VISUAIS DA FERRAMENTA DE PERFILAMENTO DE DADOS DO DATABRICKS

Figura A.1 – Tela de perfilamento do Databricks com estatísticas descritivas



Fonte: Databricks, 2025.

ANEXO B – EVIDÊNCIAS VISUAIS DA FERRAMENTA DE PERFILAMENTO DE DADOS DO YDATAPROFILING

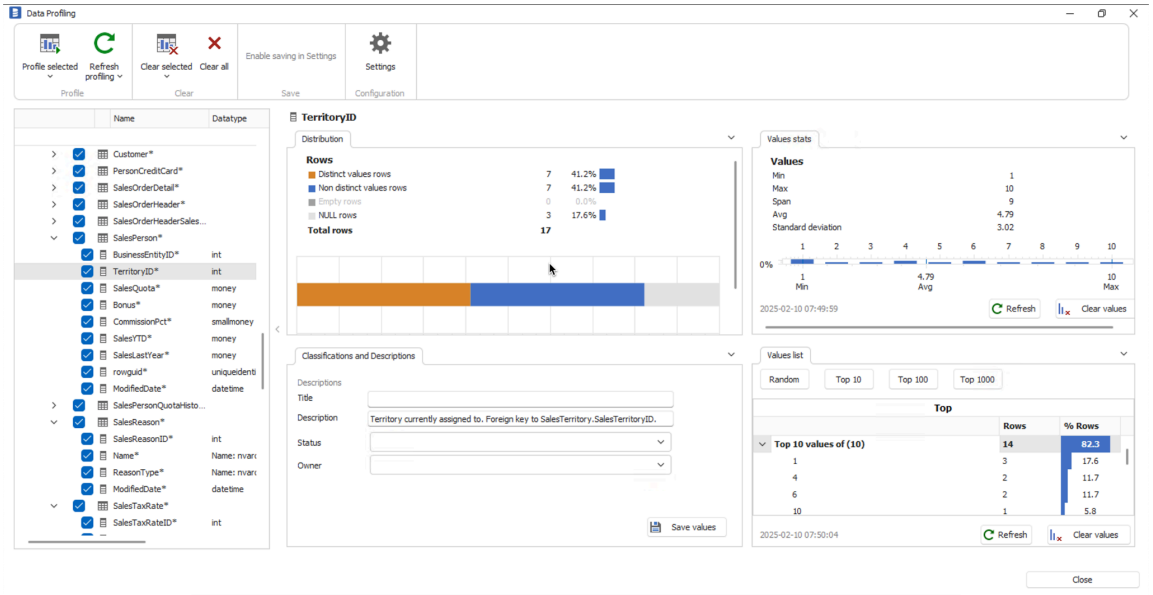
Figura B.1 – Resultado do perfilamento do yDataProfiling

Fonte: CLEMENTE et al. (2023). Relatório de perfilamento de dados. A seção Overview (1) resume as características gerais dos dados, desde o número de atributos/observações e valores duplicados/ausentes até os tipos de variáveis, o uso de memória e os problemas de qualidade encontrados nos dados. Em Variables (2), cada atributo individual pode ser investigado mais detalhadamente, verificando suas estatísticas básicas e distribuição, bem como outros indicadores avançados, como estatísticas descritivas específicas por quantis, valores comuns e valores extremos. As seções Interactions (3) e Correlations (4) oferecem suporte à análise

multivariada dos atributos, permitindo explorar as relações entre eles. Por fim, a seção Missing Values (5) fornece detalhes adicionais sobre a porcentagem de ausência de cada atributo e permite investigar os mecanismos de ausência, enquanto as seções Sample (6) e Duplicate Rows (7) fornecem uma prévia do DataFrame original e das observações duplicadas, respectivamente.

ANEXO C – EVIDÊNCIAS VISUAIS DA FERRAMENTA DE PERFILAMENTO DE DADOS DO DATAEDO

Figura C.1 – Perfilamento de dados do Dataedo



Fonte: Dataedo, 2025