UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

PAULO VITOR ALVES DE OLIVEIRA

MÉTODOS DE AMOSTRAGEM EM AMBIENTES BIG DATA PARA O DATA PROFILING: Fundamentos, Desafios e Aplicações

Recife
2025

PAULO VITOR ALVES DE OLIVEIRA

MÉTODOS DE AMOSTRAGEM EM AMBIENTES BIG DATA PARA O DATA PROFILING: Fundamentos, Desafios e Aplicações

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.

Orientador (a): Prof. Robson do Nascimento Fidalgo

Recife

2025

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

Oliveira, Paulo Vitor Alves de.
Métodos de Amostragem em Ambientes Big Data para o Data Profiling: Fundamentos, Desafios e Aplicações / Paulo Vitor Alves de Oliveira. - Recife, 2025.
63 p. : il., tab.

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

1. Big Data. 2. Perfilamento de Dados. 3. Métodos de Amostragem. 4. Qualidade de Dados. 5. Governança de Dados. 6. Revisão Sistemática da Literatura. I. Fidalgo, Robson. (Orientação). II. Título.

000 CDD (22.ed.)

PAULO VITOR ALVES DE OLIVEIRA

MÉTODOS DE AMOSTRAGEM EM AMBIENTES BIG DATA PARA O DATA PROFILING: Fundamentos, Desafios e Aplicações

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

Primeiramente, expresso minha gratidão à minha mãe, Elizabeth, por seu apoio, amor e suporte em todas as minhas decisões e jornadas, sejam elas acadêmicas, profissionais ou pessoais, sendo não só minha mãe, mas uma melhor amiga e melhor confeiteira.

Agradeço também ao meu pai, Josias, e ao meu irmão, João, que sempre estiveram ao meu lado. Sempre caminhando juntos, crescendo e aprendendo a cada dia, impulsionando a trazer uma nova e melhor versão de mim a cada dia, além de inspirar ideias, discussões e pensamentos críticos para um futuro melhor.

Um agradecimento especial ao meu orientador, Robson, que me guiou e acompanhou nesta longa jornada, atuando tanto como professor quanto mentor. Estendo minha gratidão a tantos outros professores do CIn e da UFPE que me acompanharam nesta caminhada e em minhas escolhas acadêmicas.

Sou profundamente grato aos grupos que formei durante esta trajetória, especialmente aos meus companheiros Marçal, Rodrigo, Isabelle, Diego, Júlia, Zé, LP, GM, Wil, Chico, Lipeira, Thiago e Caio. Fomos mais do que um grupo de estudo, nos ajudamos, apoiamos, criticamos e evoluímos juntos até o fim. Isso me possibilitou chegar até aqui como a pessoa que sou hoje e, com certeza, tornou minha passagem pela universidade mais divertida, vívida e prazerosa.

Agradeço também a todos que acreditaram em mim, me apoiaram, e que estiveram presentes em minha vida: Luan, Marilson, Gustavo, Helder, Mateus, Daniel, Natália e Bia. Vocês fizeram parte da minha trajetória e, por isso, merecem crédito nesta conquista.

Finalmente, expresso minha gratidão às minhas experiências profissionais e a todos os amigos que fiz nas diversas empresas por onde passei. Um agradecimento especial aos meus mentores, Vitor, Matheus e Cláudio, que me deram inúmeras sugestões de crescimento tanto no âmbito profissional quanto pessoal. Sou também grato a Caio, Bruna, Fernando, Felipe, Antonio, Esther, Rebeca e Renan, que sempre me ajudaram quando necessário e estavam lá para trazer o máximo de potencial possível.

RESUMO

O Big Data transformou-se em um pilar para a inovação e a tomada de decisão em múltiplos setores, contudo, a eficácia dessas decisões está intrinsecamente ligada à qualidade dos dados subjacentes, tornando a sua gestão um desafio de alta complexidade. A baixa qualidade dos dados pode levar a análises imprecisas, conclusões enganosas e, consequentemente, a perdas de confiança e valor. Neste cenário, o Data Profiling emergiu como um processo fundamental, atuando como uma linha de defesa e descoberta no âmbito da governança de dados e da garantia da qualidade. No entanto, o volume, a velocidade e a variedade dos dados em ambientes de Big Data tornam os métodos de perfilamento tradicionais ineficazes e computacionalmente caros. Para contornar essas limitações, as técnicas de amostragem emergem como uma solução crucial, permitindo a análise de subconjuntos representativos de dados para inferir características do conjunto total, reduzindo custos e acelerando o processamento sem comprometer significativamente a precisão. O presente trabalho constitui-se como uma Revisão Sistemática da Literatura (RSL) com o objetivo de consolidar e analisar o estado da arte sobre técnicas de amostragem em ambientes Big Data. Baseando-se em um corpus de 14 artigos científicos e publicações técnicas, esta revisão explora os fundamentos conceituais das técnicas de amostragem, detalhando suas definições, propósitos e vantagens de utilização. Por fim, a revisão conclui com uma síntese dos achados e propõe direções para pesquisas futuras, destacando a trajetória da amostragem em ambientes Big Data como uma técnica cada vez mais inteligente, escalável e consciente do contexto, indispensável para extrair valor confiável na era do Big Data, especialmente para o Data Profiling.

Palavras-chave: Perfilamento de Dados; Big Data; Métodos de Amostragem; Qualidade de Dados; Revisão Sistemática da Literatura

ABSTRACT

Big Data has become a foundational pillar for innovation and data-driven decision-making across various sectors. However, the effectiveness of such decisions is intrinsically linked to the quality of the underlying data, making data management a highly complex challenge. Low data quality can result in inaccurate analyses, misleading conclusions, and consequently, loss of trust and value. In this context, data profiling emerges as a fundamental process, serving both as a diagnostic and discovery tool within data governance and quality assurance frameworks. Nonetheless, the volume, velocity, and variety of data in Big Data environments render traditional profiling methods inefficient and computationally expensive. Sampling techniques offer a critical solution to this limitation by enabling the analysis of representative data subsets. These techniques reduce processing time and costs while maintaining acceptable levels of accuracy, particularly in large-scale scenarios. This study presents a Systematic Literature Review (SLR) that consolidates and analyzes the state of the art in sampling techniques for Big Data environments. Based on a corpus of 14 peer-reviewed scientific and technical publications, the review explores the conceptual foundations of sampling, outlining its definitions, purposes, and advantages. The work concludes with a synthesis of key findings and proposes directions for future research, emphasizing the role of sampling as an increasingly intelligent, scalable, and context-aware technique—one that is indispensable for extracting trustworthy insights from Big Data, particularly in data profiling tasks.

Keywords: Data Profiling; Big Data; Sampling Methods; Data Quality; Systematic Literature Review

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CQ	Critério de Qualidade
AI	Artificial Intelligence
IA	Inteligência Artificial
RQ	Research Question
HDFS	Hadoop Distributed File System
RSL	Revisão Sistemática da Literatura

SUMÁRIO

1 INTRODUÇÃO

O avanço exponencial da tecnologia e da ciência da computação nos últimos anos tem sido um motor fundamental para a geração de um volume sem precedentes de informações digitais. Esse fenômeno, agora amplamente reconhecido como Big Data, não apenas redefine a maneira como interagimos com o mundo digital, mas também inaugura um cenário complexo e multifacetado, repleto de novas oportunidades e, simultaneamente, de desafios significativos.

Nesse contexto, a qualidade dos dados é um fator crucial para as empresas, especialmente no que tange à tomada de decisões ágeis e automatizadas. No entanto, embora a demanda por big data esteja aumentando, ainda faltam pesquisas sistemáticas sobre como utilizá-los, melhores oportunidades e formas de integrar no escopo de Big Data. (ABUDUZZAMAN; HASAN, 2022)

Essa realidade impulsiona a necessidade de desenvolver métodos para perfilar, caracterizar e categorizar automaticamente a qualidade desses dados. No entanto, é um desafio significativo para alguns algoritmos tradicionais de mineração de dados, aprendizado de máquina e, especialmente, tarefas de Data Profiling, lidar com uma quantidade tão vasta de informações. Assim, diante da escala esmagadora dos conjuntos de dados modernos, os profissionais da ciência de dados têm duas opções estratégicas principais: adaptar algoritmos analíticos existentes para funcionar em ambientes distribuídos de larga escala ou reduzir o tamanho dos dados para um subconjunto gerenciável que pode ser processado com as ferramentas existentes. (DJOUZI; BEGHDAD-BEY; AMAMRA, 2022)

O presente trabalho concentra-se na última estratégia, centrada nos princípios de amostragem de dados. Assim, para atingir este objetivo, conduziu-se uma revisão sistemática da literatura, assegurando um rigoroso processo de seleção e análise das publicações mais recentes e pertinentes sobre o tema. Este estudo visa também identificar as lacunas na literatura atual e sugerir direções para futuras investigações. Desse modo, espera-se oferecer uma contribuição tanto para o desenvolvimento teórico quanto para o suporte prático na implementação dessa abordagem por profissionais e organizações.

A estrutura desta monografia é a seguinte: o Capítulo 2 aborda os conceitos essenciais de Data Profiling e amostragem, que são cruciais para as perguntas de pesquisa. A metodologia, com ênfase na Revisão Sistemática da Literatura, é

detalhada no Capítulo 3. Em seguida, o Capítulo 4 apresenta e discute os resultados obtidos, em relação às perguntas de pesquisa. No Capítulo 5, é realizada uma análise crítica desses resultados, destacando suas implicações teóricas e práticas. As ameaças à validade do estudo são discutidas no Capítulo 6. Por fim, o Capítulo 7 apresenta as conclusões, contribuições, limitações e sugere pesquisas futuras.

2 CONCEITOS IMPORTANTES

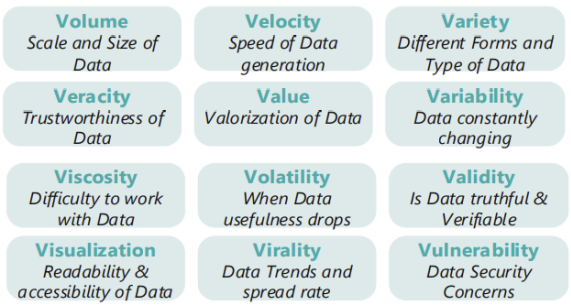
2.1 Big Data

Big Data refere-se a conjuntos de dados cujo tamanho e complexidade tornam seu processamento um desafio para as ferramentas de gerenciamento de banco de dados e aplicações de processamento de dados tradicionais. A definição vai além do mero tamanho, abrangendo um conjunto de características multifacetadas, comumente conhecidas como os "Vs" do Big Data. As três características mais consolidadas, propostas inicialmente pela Gartner, são Volume, Velocidade e Variedade.

No entanto, Big Data é algo grande e complexo que é difícil ou impossível para sistemas e ferramentas tradicionais processá-lo e trabalhar nele. A IBM o descreve através do modelo "5 Vs": Volume, que se refere à vasta quantidade de dados, representando o principal desafio para sistemas convencionais; Velocidade, indicando a rapidez com que os dados são gerados; Variedade, que engloba a diversidade de fontes e tipos de dados, como estruturados, semiestruturados e não estruturados; Valor, a característica mais crucial do Big Data; e Veracidade, que aborda a qualidade dos dados, frequentemente com a presença de dados "sujos". Devido à imensidão do Big Data, sua análise e mineração demandam alto poder computacional e capacidade de armazenamento. Além disso, algoritmos de mineração clássicos podem levar horas ou dias para produzir resultados, exigindo múltiplas passagens por todo o conjunto de dados (LIU; ZHANG, 2020).

Embora essas sejam as principais definições, outros autores estendem essa definição para abranger outras 9 características extras do Big Data que foram sendo descobertas ao longo dos anos. (TALEB; SERHANI; DSSOULI, 2019)

Figura 1 – Definição de Big Data



Fonte: (TALEB; SERHANI; DSSOULI, 2019)

Essa complexidade ressalta que o valor do Big Data não reside em sua mera posse, mas sim na análise. Contudo, a confiabilidade dessa análise depende intrinsecamente da qualidade dos dados, pois a incerteza e a potencial falta de confiabilidade dos dados brutos são os principais motivadores para que as organizações implementem estruturas robustas de validação, limpeza e gerenciamento desses ativos informacionais em larga escala, como o Data Profiling e as técnicas de amostragem, que serão abordados nos próximos capítulos.

2.2 Data Profiling

Data Profiling é fundamentalmente o processo de examinar dados para desenvolver uma compreensão abrangente de suas características, estrutura, conteúdo e qualidade, principalmente por meio da geração de metadados e resumos informativos, ou seja, é um processo investigativo que é crucial para transformar dados brutos em um ativo bem compreendido. Diversos pesquisadores propuseram definições que, embora variem em suas formulações específicas, convergem para este tema central da descoberta de metadados e compreensão de dados.

Abiduzzaman e Hasan (2022) definem Data Profiling como o processo de estudar, analisar e desenvolver resumos relevantes de dados, que fornece uma "visão geral de alto nível" crucial para identificar preocupações com a qualidade dos dados, riscos potenciais e tendências abrangentes. Eles ainda explicam que esse processo examina os dados para verificar sua legalidade e qualidade, empregando algoritmos analíticos para detectar diversas propriedades, como média, valores mínimos/máximos, percentis e distribuições de frequência. Isso, por sua vez, revela metadados essenciais, incluindo distribuições de frequência e relacionamentos-chave dentro dos dados.

Da mesma forma, Elbaghazaoui et al. (2021) descrevem o Data Profiling como o processo de descoberta de metadados por meio do meticuloso exame dos conjuntos de dados e coleta de metadados, como estatísticas ou resumos informativos sobre esses dados. Já Liu e Zhang (2020) oferecem uma definição concisa, afirmando que é a atividade que encontra metadados do conjunto de dados.

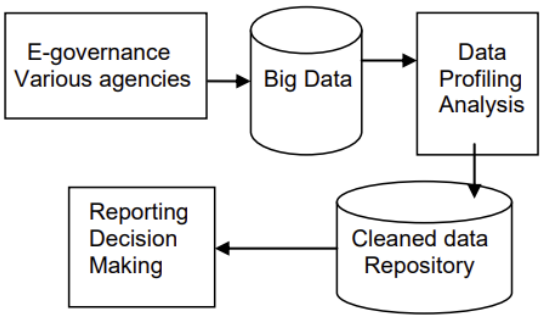
Uma definição mais abrangente é fornecida por Dai et al. (2016), que caracteriza Data Profiling como o processo de verificação de dados estruturados, semiestruturados e não estruturados dos usuários, coletando estrutura de dados, padrões de dados, informações estatísticas, mensagens de distribuição e revisão de

atributos de dados para governança, gerenciamento, migração e controle de qualidade de dados. Essa definição reconhece explicitamente as diversas formas que os dados podem assumir em ambientes modernos. Juddoo (2015) também contribui ao defini-la como o exame de fontes de dados para gerar informações sobre os dados e conjuntos de dados, incluindo estatísticas e metadados, relacionamentos, dependências, padrões e cardinalidades.

A convergência dessas definições sobre descoberta de metadados e compreensão de dados ressalta o objetivo principal do Data Profiling. A sutil evolução nessas definições, particularmente a inclusão explícita de dados semiestruturados e não estruturados por Dai et al. (2016), reflete a crescente complexidade e variedade dos cenários de dados contemporâneos. As primeiras conceituações de Data Profiling podem ter se centrado implicitamente em dados estruturados e relacionais. No entanto, à medida que as fontes de dados se diversificaram, uma característica fundamental da era do Big Data, o escopo do Data Profiling necessariamente se expandiu.

Essa expansão implica que as técnicas e ferramentas utilizadas para o Data Profiling também devem evoluir para lidar efetivamente com essa diversidade, indo além das abordagens tradicionais projetadas principalmente para bancos de dados relacionais. A adaptação do campo a essas novas realidades é crucial para manter a relevância e a eficácia do Data Profiling diante dos paradigmas de dados em constante mudança.

Figura 2 – Etapas de Data Profiling



Fonte: (BABU; KUMAR, 2019)

2.2.1 Qualidade e Governança de Dados

Qualidade e Governança de Dados são conceitos interligados ao Data Profiling, uma vez que todos são essenciais na gestão de dados. Juntos, garantem o

uso adequado, a eficiência e a precisão das aplicações. Abiduzzaman & Hasan (2022) afirmam que o Data Profiling auxilia na descoberta de preocupações, riscos e tendências gerais em relação à qualidade dos dados, prevenindo resultados analíticos imprecisos. Já Shivaprasad (2024) reforça que o Data Profiling automatizado melhora a tomada de decisão, sendo uma solução popular para automatizar a precisão e a qualidade dos dados. Temos também Rangineni et al. (2023) que destacam sua importância na identificação e correção de erros e Azeroual et al. (2018) que comprovam sua aplicabilidade na pesquisa de sistemas de informação.

Essencialmente, o Data Profiling funciona como um diagnóstico que identifica problemas como valores ausentes, inconsistências e anomalias, conforme dito por Abiduzzaman & Hasan (2022) e Shivaprasad (2024), além de orientar as ações corretivas de melhoria da qualidade dos dados, como limpeza e transformação, sugeridas por Rangineni et al. (2023), garantindo intervenções eficazes.

Sua importância é enorme, uma vez que a qualidade de dados é uma medida da condição dos dados com base em fatores como precisão, integridade, consistência, atualidade, exclusividade e relevância para uma finalidade específica. Dados de baixa qualidade podem levar a análises defeituosas, decisões equivocadas e perdas financeiras e de reputação significativas. Por exemplo, Abiduzzaman & Hasan (2022) cita estimativas de que dados imprecisos podem reduzir os lucros em até 30%, resultando em perda de produtividade e oportunidades de vendas.

Figura 3 – Dimensões da Qualidade de Dados

Fonte: (RANGINENI et al., 2023)

Por outro lado, a Governança de Dados é o controle e a tomada de decisão sobre os ativos de dados de uma organização, definindo políticas e responsabilidades para garantir a segurança, consistência e conformidade dos dados. Dai et al. (2016) e Babu & Kumar (2019) destacam o valor do perfilamento para a governança e gerenciamento de dados mestres e Maddali (2023) sugere que insights de IA, gerados pelo Data Profiling, podem automatizar a governança.

Em essência, o Data Profiling oferece a base factual para as estruturas de governança, informando regras de padronização e atribuição de responsabilidades, sendo um componente dinâmico para a governança de dados de próxima geração. Sem ele, as iniciativas de governança seriam abstratas e ineficazes.

2.2.3 Tarefas do Data Profiling

A literatura acadêmica e técnica converge consistentemente para uma taxonomia de tarefas de Data Profiling organizada em três níveis de complexidade crescente, e esta classificação fornece uma estrutura lógica para entender as diferentes facetas da análise de dados. Elas podem ser classificadas de acordo com o tipo de dados que processam, sendo divididas em: perfilamento de coluna única, perfilamento de multicolunas e perfilamento de dependências. (ELBAGHAZAOUI; AMNAI; SEMMOURI, 2021)

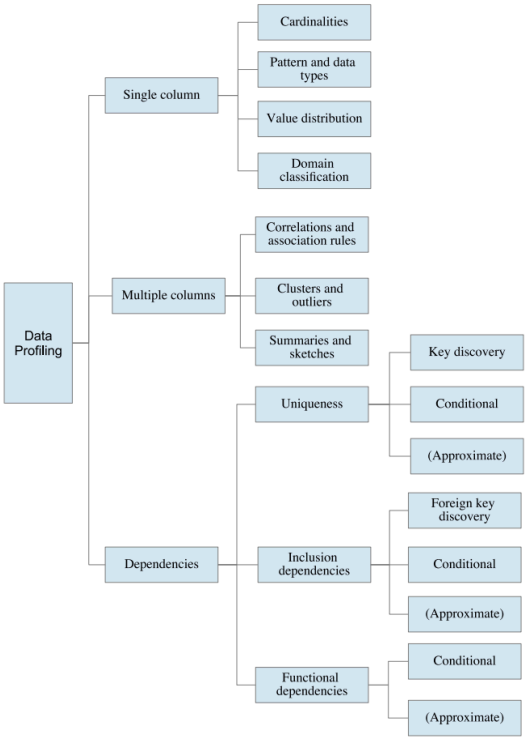
O perfilamento de coluna única é a forma mais básica de perfilamento, focada em analisar os valores dentro de uma única coluna, assumindo que todos os valores compartilham o mesmo tipo e propriedades comuns. As tarefas incluem: estatísticas descritivas, com a coleta de métricas como mínimo, máximo, média, soma e contagem; cardinalidade, com a contagem do número de valores distintos em uma coluna, incluindo a contagem de nulos e a porcentagem de valores únicos; distribuição de valores, com a frequência com que cada valor aparece, o que pode revelar a distribuição dos dados; e padrões e tipos de dados, com a identificação de padrões recorrentes e inferência do tipo de dados.

Já o perfilamento de múltiplas colunas avança para descobrir propriedades conjuntas e dependências entre diferentes colunas dentro da mesma tabela. A complexidade aqui aumenta significativamente, pois envolve a comparação de combinações de colunas. As principais tarefas são: correlações e regras de associação, com a utilização de métodos estatísticos para descobrir como os valores em diferentes colunas se influenciam mutuamente, revelando associações; e a

detecção de clusters e outliers, com a identificação de subgrupos de registros com características semelhantes e detecção de pontos de dados que se desviam significativamente do padrão.

Por fim, o perfilamento na descoberta de dependências é o nível mais complexo e computacionalmente intensivo, focado na descoberta de regras estruturais e de integridade que governam os dados. As tarefas incluem: dependências funcionais, com descoberta de regras onde o valor de uma coluna determina univocamente o valor de outra; dependências de inclusão, com a verificação de que o conjunto de valores de uma coluna está contido no conjunto de valores de outra, geralmente em uma tabela diferente; e combinações de colunas únicas, com a identificação de conjuntos de colunas cujos valores combinados são únicos para cada registro na tabela. Nessa etapa é onde temos as ações que nos ajudam principalmente a identificar as chaves primárias candidatas e as chaves estrangeiras candidatas.

Figura 4 – Tarefas de Data Profiling



Fonte: (LIU; ZHANG, 2020)

2.3 Sampling

A amostragem é formalmente definida como um procedimento de análise estatística usado para selecionar, gerenciar e analisar um subconjunto representativo de dados (DJOUZI; BEGHDAD-BEY; AMAMRA, 2023). De forma mais fundamental, é um método científico de seleção de dados de amostra representativos da base de dados, ou seja, consiste em usar menos dados para obter as características gerais de todo o conjunto de dados (LIU; ZHANG, 2020).

O objetivo principal é reduzir efetivamente a quantidade de dados e ajudar a acelerar o processamento de dados, mitigando assim as cargas computacionais e de recursos associadas ao Big Data. A proposta de valor central da amostragem está em sua capacidade de gerar resultados de um subconjunto pequeno e bem escolhido que são estatisticamente próximos ou até mesmo excedem os resultados da quantidade total de dados. Isso permite que os analistas alcancem um equilíbrio entre eficiência computacional e precisão analítica, tornando possível derivar insights oportunos sem incorrer nos custos proibitivos de análises exaustivas. (LIU; ZHANG, 2020)

Um olhar atento da literatura revela que o termo "amostragem" abrange um amplo espectro de técnicas que atendem a dois propósitos estratégicos distintos e, por vezes, conflitantes. O primeiro, enraizado na estatística clássica, é a busca pela representatividade, o objetivo aqui é criar uma amostra que seja um subconjunto representativo reduzido de toda a população. Como afirmam Mahmud et al. (2020), o objetivo é obter pequenos subconjuntos para generalizar com precisão os resultados da amostra para todo o conjunto de dados, essa abordagem é essencial para tarefas que exigem estimativas globais precisas. O segundo objetivo é a exploração e a descoberta direcionada, isso envolve a criação de uma amostra intencionalmente tendenciosa ou focada para acelerar uma tarefa analítica específica.

Isso revela uma dualidade fundamental na estratégia de amostragem: a escolha de uma técnica é uma decisão estratégica que depende se o objetivo final é um perfil global preciso, que exige representação, ou uma análise eficiente e direcionada, que pode se beneficiar do viés estratégico.

Como Taleb (2016) explicitamente observa, avaliar a qualidade do Big Data, uma função central do Data Profiling, é um processo custoso se aplicado a todos os dados. Essa inviabilidade computacional e econômica força uma mudança crítica de paradigma, afastando-se da busca por análises exatas e completas, uma vez que as características definidoras do Big Data tornam tal objetivo insustentável para muitas aplicações do mundo real.

Consequentemente, o campo deve transitar da computação exata para a aproximação inteligente. Essa mudança não é uma questão de conveniência, mas sim uma resposta a restrições fundamentais. Assim, torna-se essencial compreender o conceito de amostragem, quais são as técnicas disponíveis e como podemos utilizá-las, sendo o principal objetivo deste trabalho trazer justamente a resposta para essas perguntas, como serão apresentadas nas próximas seções.

2.3.1 Record Level Sampling e Block Based Sampling

Amostragem em nível de registro é a abordagem tradicional, em que o algoritmo de amostragem examina e seleciona pontos de dados individuais, como linhas em uma tabela, entradas de log únicas ou documentos individuais, um por um. Todos os métodos clássicos descritos na seção 4.1, em suas formas ingênuas, são operações em nível de registro.

Já a amostragem em nível de bloco é uma abordagem mais recente, que surgiu da arquitetura de sistemas de arquivos distribuídos, como o Hadoop Distributed File System (HDFS). Nesses sistemas, arquivos grandes são divididos em blocos fixos, por exemplo, 128 MB ou 256 MB, que são então distribuídos entre as máquinas de um cluster. Essa metodologia de amostragem considera esses blocos de dados inteiros, e não registros individuais, como as unidades básicas a serem amostradas.

Conforme observado por Mahmud et al. (2020) e Liu & Zhang (2020), enquanto a amostragem por registro exige muita comunicação de rede e de disco para recuperar dados espalhados, a amostragem por bloco é mais eficiente, recuperando blocos inteiros com menor sobrecarga.

3. Metodologia

Este estudo foi conduzido seguindo o modelo metodológico de Kitchenham e Charters (2007), que organiza o processo de mapeamento em três fases principais: planejamento, execução e apresentação dos resultados. O planejamento, fase inicial, foca na identificação da necessidade do mapeamento e na formulação das questões de pesquisa. A execução, segunda fase, abrange a seleção de estudos por meio da identificação, triagem, avaliação de qualidade, extração e síntese de informações. Por fim, o relato dos resultados organiza e apresenta as conclusões. A Figura 5 detalha o fluxo metodológico.

Figura 5 – Estrutura baseada em Kitchenham e Charters

Fonte: O autor (2025)

Adotou-se a metodologia de Kitchenham e Charters (2007) devido à sua solidez em Ciência de Dados e Engenharia de Software, além de que essa abordagem oferece um método estruturado para a identificação, seleção e análise de estudos, garantindo uma coleta e síntese imparciais, junto de critérios de qualidade para refinamento. Por meio dessa abordagem, o presente trabalho visa mapear os fundamentos e benefícios das técnicas de amostragem em Big Data, consolidando práticas emergentes e fornecendo subsídios para futuras pesquisas e implementações no contexto organizacional.

3.1 Definição das questões de Pesquisa

O objetivo do estudo é analisar as técnicas e métodos de amostragem aplicados no contexto de Big Data, buscando identificar a possibilidade de sua utilização em Data Profiling. Contudo, o foco principal não é o Data Profiling em si, mas sim apresentar essas técnicas que podem ser úteis nesse contexto. Assim, foi estruturada por três questões centrais: a RQ1 explora quais são as técnicas e métodos de amostragem; a RQ2 analisa e mapeia quais são os pontos fortes e fracos de cada amostragem; e a RQ3 analisa e mapeia quais são e onde são suas principais aplicações.

Essa abordagem possibilita a consolidação de um panorama abrangente sobre as tendências, os benefícios e os desafios da amostragem em cenários de Big Data. Assim, auxilia tanto pesquisadores quanto profissionais da área a compreenderem os impactos dessas técnicas e como elas poderiam ser benéficas para o Data Profiling. As questões e suas respectivas motivações estão detalhadas na Tabela 1.

Tabela 1 – Questões de Pesquisa

Questões de Pesquisa	Motivações
RQ1. Quais são as principais técnicas de amostragem encontradas na literatura para Big Data?	Analisar as principais atuações recentes, buscando uma compreensão aprofundada.
RQ2. Quais são os pontos fortes e pontos fracos de cada técnica de amostragem?	Analisar os benefícios e as barreiras enfrentadas por cada técnica apresentada.
RQ3. Quais são as principais aplicações que se beneficiam dessas técnicas?	Analisar os escopos de aplicação e as principais recomendações de cada técnica.

Fonte: O autor (2025).

3.2 Estratégia de Busca

A fim de identificar estudos relevantes sobre técnicas de amostragem, foram definidas palavras-chave alinhadas às questões de pesquisa. A seleção dos termos levou em consideração os conceitos técnicos de amostragem, com foco no contexto de Big Data e suas potenciais ligações com aplicações em Data Profiling. A string de busca foi refinada por meio de testes iniciais em bases acadêmicas, utilizando operadores booleanos para otimizar a recuperação de publicações dentro do escopo do mapeamento. A formulação final adotada foi:

("sampling techniques" OR "sampling methods" or “sampling”) AND ("big data" or "large-scale data" or "profiling")

A pesquisa automatizada foi conduzida por meio do Google Scholar, uma plataforma amplamente reconhecida por sua acessibilidade e vasta coleção de artigos acadêmicos relevantes. O Google Scholar atua como um agregador de conteúdo de fácil acesso, integrando diversas bases de dados como IEEE, ACM e Scopus, sendo o principal motivo da escolha.

3.3 Critérios de Inclusão e Exclusão

Esta pesquisa concentra-se exclusivamente em estudos diretamente relacionados a métodos e técnicas de amostragem no ambiente de Big Data. Para assegurar a qualidade e relevância das publicações, foram estabelecidos critérios rigorosos de inclusão e exclusão.

Priorizamos artigos completos de 2015 a 2025, de periódicos científicos e anais de conferências relevantes em dados e computação. Selecionamos estudos que descrevessem técnicas, desafios ou ferramentas de amostragem em Big Data e apresentassem casos reais ou avaliações práticas para que seja possível o entendimento e a continuidade da revisão sistemática. Além disso, todos os artigos selecionados estão em inglês para facilitar o entendimento e padronização da língua técnica utilizada.

Excluímos estudos fora do período definido, sem acesso ao texto completo, duplicados ou que não abordassem diretamente a amostragem em Big Data, ou que fossem apenas descritivos sem análise metodológica, como por exemplo técnicas soltas ou menções rápidas sem nenhum contexto específico. Os critérios completos de inclusão e exclusão estão nas Tabelas 2 e 3, respectivamente.

Tabela 2 – Critérios de Inclusão

#	Critérios de Inclusão
C1	Artigos que descrevem técnicas, desafios ou ferramentas de amostragem no ambiente de Big Data
C2	Estudos que apresentem casos reais ou avaliações práticas sobre ferramentas de amostragem em ambientes Big Data
C3	Artigos em inglês
C4	Publicações em periódicos e conferências relevantes da área
C5	Estudos publicados entre 2015 e 2025

Fonte: O autor (2025).

Tabela 3 – Critérios de Exclusão

#	Critérios de Exclusão
C1	Estudos fora do período definido (anteriores a 2015)
C2	Artigos sem acesso ao texto completo
C3	Trabalhos que não abordam diretamente o tema de amostragem em contexto de Big Data
C4	Trabalhos duplicados nas bases
C5	Artigos apenas descritivos sem análise metodológica

Fonte: O autor (2025).

Os critérios de idioma adotados nesta revisão sistemática foram definidos para assegurar o pleno entendimento do conteúdo dos artigos selecionados, evitando potenciais limitações decorrentes de traduções que poderiam comprometer a acurácia na análise e interpretação dos dados sobre amostragem em Big Data. Ao restringir a busca a artigos em inglês, garantimos uma terminologia padronizada na área, facilitando entendimento, busca e mapeamento da revisão sistemática.

Já o período de tempo buscou avaliar tanto o crescimento recente de pesquisas na área de amostragem em Big Data quanto evitar a análise de metodologias ou ferramentas mais antigas que poderiam não refletir a eficácia e os desafios dos modelos e tecnologias atuais. Isso torna o estudo mais coerente e relevante, focado nos avanços mais recentes e nas práticas emergentes de amostragem em ambientes de Big Data. Apesar do período ser mais extenso do que o usual, o objetivo é mapear as antigas formas de atuação, as tendências em ascensão e as novas abordagens recentes, mas sempre se atentando ao risco de

incluir pesquisas defasadas e evitar de tornar a análise heterogênea em relação ao estado da arte.

3.4 Critérios de Qualidade

Para garantir o rigor metodológico e a relevância científica dos estudos selecionados sobre métodos e técnicas de amostragem em ambientes de Big Data, foi implementada uma avaliação de qualidade baseada em critérios estruturados. A aplicação desses critérios em revisões sistemáticas é fundamental, conforme destacado por Kitchenham e Charters (2007), para mitigar vieses, aprimorar a validade interna e externa e assegurar a confiabilidade dos resultados.

Os critérios de qualidade englobam fundamentação teórica, metodologia, apresentação de resultados e relevância científica, além disso cada critério foi pontuado, permitindo uma análise comparativa eficaz entre as publicações assegurando a seleção de estudos de alta qualidade e confiáveis sobre os fundamentos, desafios e tendências das técnicas de amostragem. A Tabela 4 detalha os critérios avaliativos.

Tabela 4 – Critérios de Avaliação de Qualidade

#	Pergunta	Resposta
CQ 1	A pesquisa oferece uma justificativa explícita para sua execução?	S=1, N=0, P=0.5
CQ 2	A definição dos objetivos da pesquisa está clara?	S=1, N=0, P=0.5
CQ 3	O ambiente no qual o estudo foi conduzido é apresentado de maneira aprofundada?	S=1, N=0, P=0.5
CQ 4	Os resultados foram apresentados de forma clara e bem estruturada?	S=1, N=0, P=0.5
CQ 5	A aplicabilidade dos achados foi avaliada na pesquisa?	S=1, N=0, P=0.5
CQ 6	O estudo aborda suas limitações metodológicas e as potenciais ameaças à validade dos resultados?	S=1, N=0, P=0.5

Fonte: O autor (2025).

3.5 Seleção de Estudos

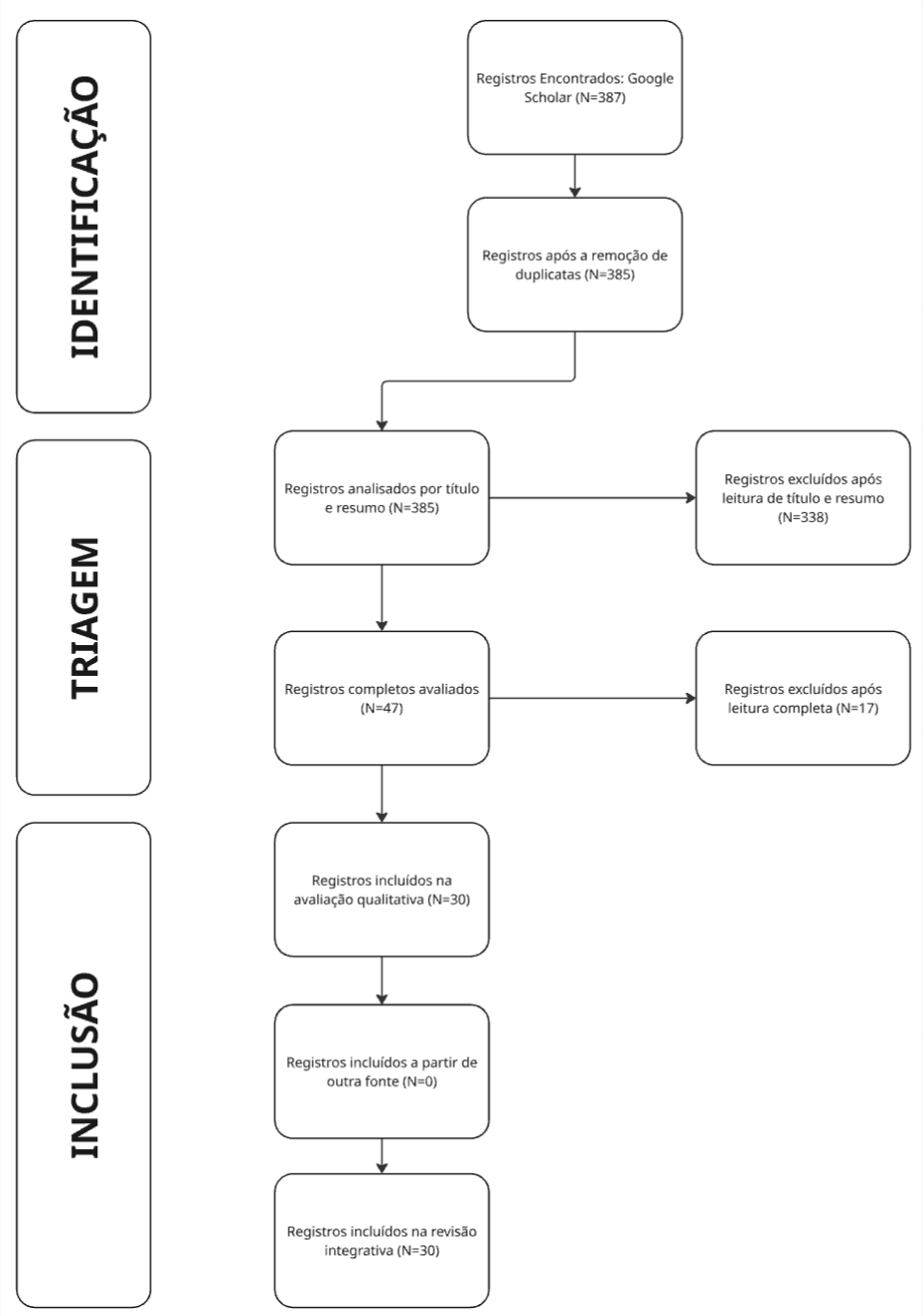
A seleção dos estudos foi realizada por meio de uma triagem rigorosa, aplicando critérios de exclusão predefinidos para alinhar os artigos aos objetivos da pesquisa. Das 401 publicações encontradas nas bases de dados, muitas foram descartadas por não se adequarem aos critérios estabelecidos.

Inicialmente, dos artigos identificados, 387 se enquadraram no período de 2015 a 2025. Posteriormente, esse número foi reduzido para 372 devido à falta de acesso completo a alguns artigos. A fase mais rigorosa de seleção foi a avaliação da relevância temática, onde 338 artigos foram excluídos por não abordarem diretamente a amostragem em Big Data com aplicações e explicações explícitas. Muitos desses trabalhos cobriam áreas correlatas, como inteligência artificial genérica, medicina ou segurança de dados, mas sem o foco necessário em amostragem, resultando em 34 artigos finais.

Por fim, 4 artigos foram removidos por serem apenas descritivos e carecerem de análise metodológica ou serem duplicatas. Ao final de todas as etapas de triagem, 30 artigos foram selecionados para compor a análise final, assegurando que apenas os trabalhos mais robustos, relevantes e alinhados ao escopo da pesquisa fossem incluídos. A Figura 6 ilustra detalhadamente todo o processo de seleção dos artigos realizados nesta etapa.

A lista completa dos artigos identificados, junto do registro do protocolo completo da RSL, bem como os critérios de triagem e avaliação aplicados com base na leitura de título e resumo está publicamente acessível no repositório Github. Este repositório contém os dados brutos da análise, o que promove a transparência e reprodutibilidade de futuras replicações deste mapeamento por outros pesquisadores. Adicionalmente, as etapas de qualidade detalhadas na seção 3.4 já estão incluídas também.

Figura 6 – Fluxo do processo de seleção dos artigos

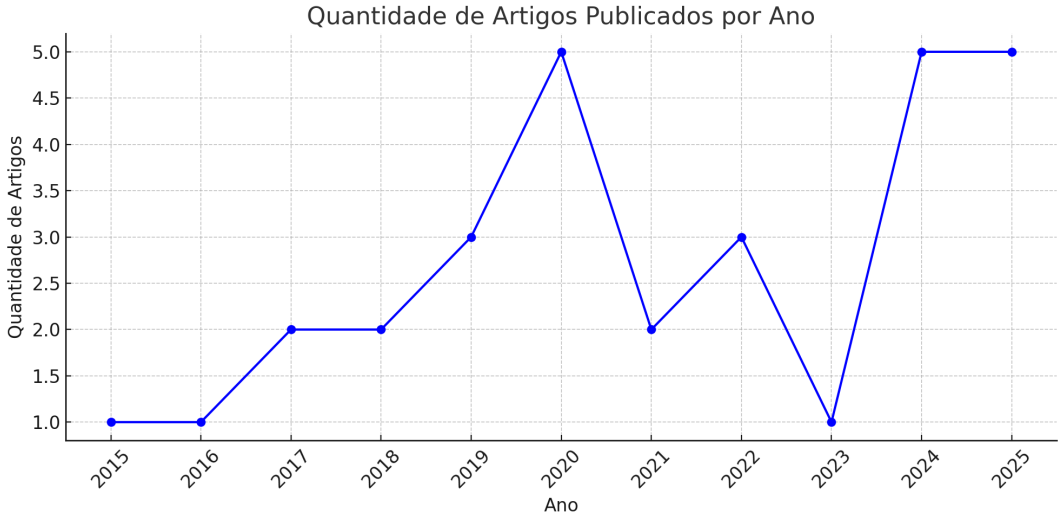


Fonte: elaborado de acordo com o modelo de Moher et al. (2009)

Após a seleção dos artigos, foi elaborada uma representação gráfica para aprofundar a compreensão da produção acadêmica e das fontes bibliográficas utilizadas. O Gráfico 1 ilustra com picos específicos, mas ao mesmo tempo um crescimento consistente no interesse por essa área. A distribuição temporal das publicações evidencia um aumento progressivo no volume de estudos voltados a técnicas de amostragem aplicadas ao contexto de Big Data, especialmente a partir

de 2020. Além disso, o maior número de publicações entre 2020 e 2025 sugere que este é um campo em expansão e alta relevância contemporânea, reforçando a atualidade da revisão sistemática, pois permite incorporar os avanços mais recentes da literatura, evita limitações comuns em revisões desatualizadas, que ignoram técnicas modernas ou tecnologias emergentes e confirma que os desafios atuais do Big Data ainda estão sendo discutidos.

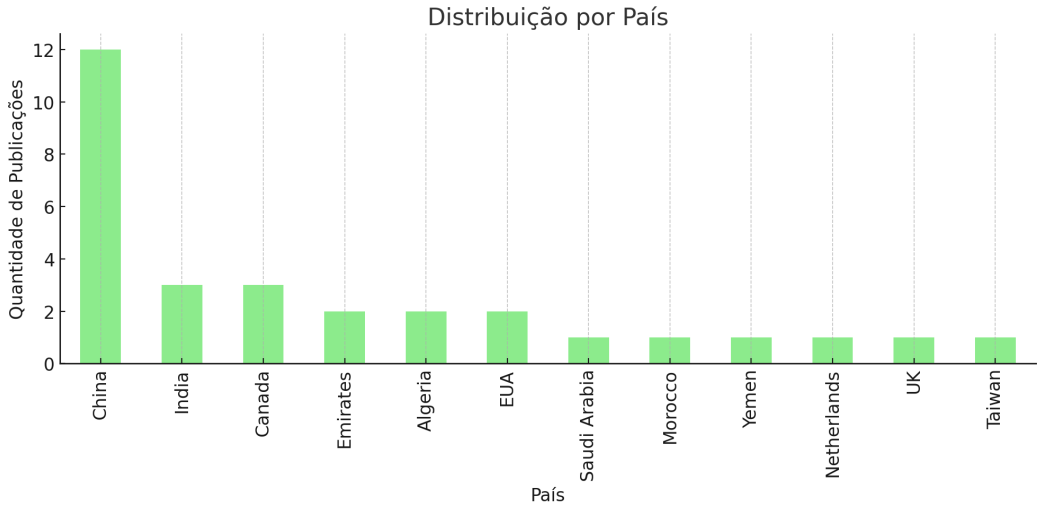
Gráfico 1 – Gráfico da quantidade de artigos publicados por ano



Fonte: O autor (2025).

Já uma análise geográfica proposta pelo Gráfico 2 revelou que, embora a China concentre a maior parte dos estudos, há participação significativa de autores de 13 países diferentes, incluindo Índia, Canadá, EUA, Argélia, Emirados Árabes, Taiwan, Reino Unido, Marrocos, Holanda, Arábia Saudita e Iêmen. Esta distribuição geográfica é importante por três motivos: evitar viés regional ou cultural, uma vez que estudos provenientes de diferentes continentes tendem a abordar problemas sob perspectivas técnicas e sociais distintas; aumentar a diversidade metodológica, pois autores de diferentes países aplicam métodos variados; e conferir maior generalização dos resultados, ao não depender exclusivamente de soluções de uma região, o trabalho se torna mais robusto e aplicável em diferentes contextos.

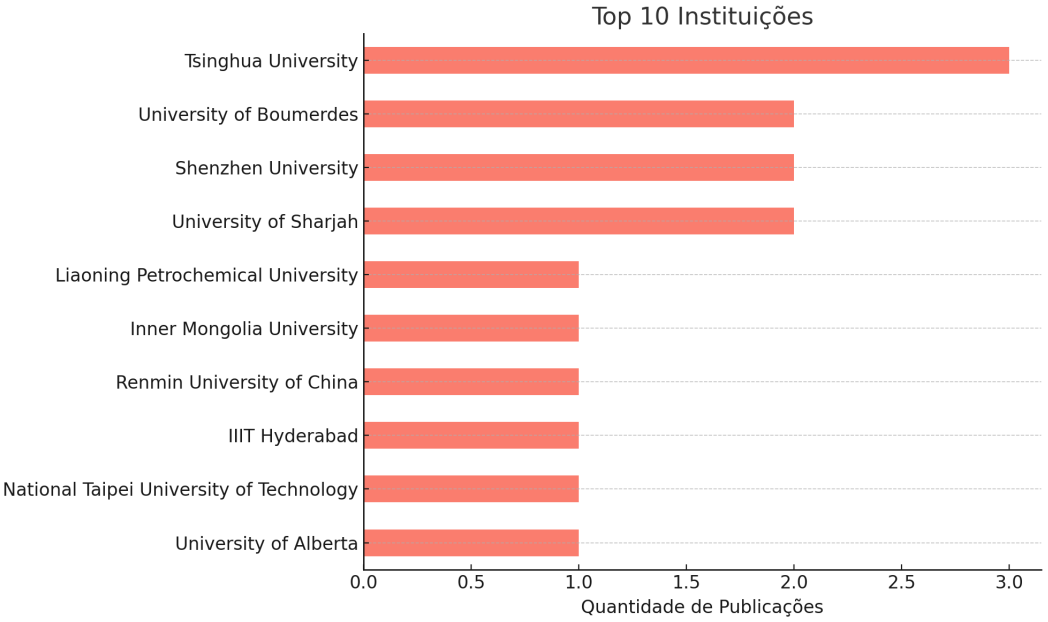
Gráfico 2 – Gráfico da quantidade de artigos publicados por país



Fonte: O autor (2025).

Uma análise das instituições apresentada no Gráfico 3 mostra que, embora algumas universidades como Tsinghua University e Shenzhen University apareçam em mais de um estudo, a amostra inclui mais de 25 instituições diferentes, com foco em pesquisa, tecnologia, e engenharia de dados. Esse ponto é relevante porque reduz o viés de afiliação, já que não há dependência de uma única escola de pensamento ou grupo de pesquisa. Por fim, mostra que a área é de interesse transversal, sendo investigada por universidades tradicionais, tecnológicas e centros emergentes.

Gráfico 3 – Gráfico da quantidade de artigos publicados por instituição afiliada



Fonte: O autor (2025).

Além disso, após a etapa final desses 30 artigos, foi realizada também uma avaliação de qualidade para garantir que os estudos estivessem em conformidade com os critérios metodológicos estabelecidos para o mapeamento. O objetivo desta fase foi verificar o rigor metodológico, a clareza dos resultados e a relevância dos estudos para a investigação das técnicas de amostragem.

Tabela 5 – Distribuição dos estudos em relação aos critérios de qualidade

Artigo	CQ1	CQ2	CQ3	CQ4	CQ5	CQ6	Total
1	1	1	1	1	0.5	0.5	83.33%
2	1	1	1	1	0.5	0.5	83.33%
3	1	1	1	1	0.5	0.5	83.33%
4	1	1	0.5	0.5	1	0.5	75.00%
5	1	1	0.5	1	1	1	91.67%
6	1	1	0.5	1	1	0.5	83.33%
7	1	1	1	1	1	0.5	91.67%
8	1	1	0.5	1	1	0.5	83.33%
9	0.5	1	0.5	1	0.5	0	58.33%
10	1	0.5	0.5	1	1	0.5	75.00%
11	1	1	1	0.5	0.5	0.5	75.00%
12	1	1	1	0.5	1	0.5	83.33%
13	1	1	1	1	0.5	0.5	83.33%
14	1	1	1	1	0.5	0.5	83.33%
15	1	1	1	1	0.5	0.5	83.33%
16	1	1	1	1	0.5	1	91.67%
17	1	0.5	0.5	1	1	0.5	75.00%
18	1	0.5	1	1	0.5	0.5	75.00%
19	1	0.5	1	0.5	0.5	1	75.00%
20	1	1	0.5	1	1	0.5	83.33%
21	1	1	1	0.5	0.5	0.5	75.00%
22	1	1	1	0.5	0.5	0.5	75.00%
23	1	1	0.5	1	0.5	0.5	75.00%
24	1	1	1	0.5	0.5	0.5	75.00%
25	1	1	0.5	0.5	0.5	0.5	66.67%
26	1	1	0.5	0.5	0.5	0.5	66.67%
27	1	1	1	0.5	0.5	0.5	75.00%
28	1	1	0.5	1	0.5	0.5	75.00%
29	1	1	1	0.5	0.5	0.5	75.00%
30	1	1	1	1	0.5	0.5	83.33%

Fonte: O autor (2025).

Para a Revisão Sistemática da Literatura, foi estabelecido um critério de qualidade mínima de 80%, garantindo que ele conseguisse atuar em pelo menos 5 dos 6 critérios de qualidade. Isso confirma a robustez teórica, a metodologia consistente e a relevância prática dos estudos, justificando sua inclusão na análise final. A Tabela 6 apresenta os resultados da avaliação de qualidade, listando os artigos selecionados com informações como ano de publicação, autores e título.

Tabela 6 – Síntese dos artigos/estudos analisados

Nº	Ano	Autores	Título
1	2020	Mahmud, Mohammad Sultan; Huang, Joshua Zhexue; Salloum, Salman; Emara, Tamer Z; Sadatdiynov, Kuanishbay;	A survey of data partitioning and sampling methods to support big data analysis
2	2019	Kim, Jae Kwang; Wang, Zhonglei;	Sampling techniques for big data analysis
3	2022	Djouzi, Kheyreddine; Beghdad-Bey, Kadda; Amamra, Abdenour;	A new adaptive sampling algorithm for big data classification
4	2024	Cai, Yongda; Wu, Dingming; Sun, Xudong; Wu, Siyue; Xu, Jingsheng; Huang, Joshua Zhexue;	CDFRS: A scalable sampling approach for efficient big data analysis
5	2019	Salloum, Salman; Huang, Joshua Zhexue; He, Yulin;	Exploring and cleaning big data with random sample data blocks
6	2017	He, Yulin; Huang, Joshua Zhexue; Long, Hao; Wang, Qiang; Wei, Chenghao;	I-sampling: A new block-based sampling method for large-scale dataset
7	2024	Khoei, Tala Talaei; Singh, Aditi;	Data reduction in big data: a survey of methods, challenges and future directions
8	2017	Rojas, Julian A Ramos; Kery, Mary Beth; Rosenthal, Stephanie; Dey, Anind;	Sampling techniques to improve big data exploration
9	2022	Anupama, CG; Lakshmi, C;	A Comprehensive Review on Data Partitioning and Sampling Techniques for Processing Big Data
10	2023	Djouzi, Kheyreddine; Beghdad Bey, Kadda; Amamra, Abdenour;	Big Data Sampling Techniques: A State-of-the-art Survey
11	2025	Zayed, Mohammed; Ba-Alwi, Fadl Mutaher;	Big Data Formation, Reduction, and Its Impact on Sampling: A survey
12	2020	Liu, Zhicheng; Zhang, Aoqian;	A survey on sampling and profiling over big data (technical report)
13	2024	Zhu, Bailin; Wang, Hongliang; Fan, Mi;	Constructing small sample datasets with game mixed sampling and improved genetic algorithm
14	2016	Ikbal Taleb, Hadeel T. El Kassabi, Mohamed Adel Serhani, Rachida Dssouli, Chafik Bouhaddioui	Big Data Quality: A Quality Dimensions Evaluation

Fonte: O autor (2025).

Assim, após um cuidadoso refinamento dos estudos, esta revisão propõe-se a analisar os fundamentos, benefícios e aplicações emergentes das técnicas de amostragem em Big Data, sendo uma revisão com abordagem predominantemente

teórica, justificada pelo número limitado de artigos disponíveis e pelo caráter inicial do estudo. O objetivo é fornecer um mapeamento e guia básico sobre as vantagens e o uso dessas técnicas em ambientes de Big Data. Por fim, os resultados e suas implicações serão discutidos no próximo capítulo.

4. Resultados

Este capítulo apresenta os resultados mais relevantes sobre as técnicas de amostragem em ambientes de dados em larga escala, ou seja, Big Data. A análise foi realizada para responder às questões de pesquisa propostas, oferecendo uma visão consolidada do estado da arte do tema, bem como os avanços e desafios inerentes a cada técnica.

As próximas seções abordarão cada técnica de amostragem individualmente, em vez de serem organizadas por perguntas de pesquisa. O objetivo é fornecer informações completas e relevantes sobre cada técnica, respondendo às perguntas RQ1-RQ3 ao longo do texto. Este trabalho visa mapear e entender as técnicas de amostragem, dessa forma, a decisão sobre o momento de utilizá-las, seja antes ou durante o Data Profiling, e qual técnica será utilizada é de responsabilidade exclusiva do cientista de dados e não será abordada aqui.

4.1 Amostragem Probabilística

A jornada rumo à amostragem de big data começa com os métodos clássicos que constituem a base da teoria estatística. Essas técnicas, enraizadas nos princípios da probabilidade, foram reprojetadas e reinventadas para funcionar dentro das restrições e da escala dos ecossistemas de dados modernos. Esta seção explica a lógica central desses métodos fundamentais para um público não especializado, mas sempre lembrando de detalhar suas adaptações e desafios específicos no cenário de big data.

4.1.1 Simple Random Sampling (SRS)

A Amostragem Aleatória Simples, também conhecida como Simple Random Sampling (SRS), é descrita como o método mais simples e comum de amostragem. Seu princípio fundamental é que cada item de dados na população tem uma probabilidade igual e independente de ser selecionado para a amostra, como se fosse retirar nomes de cada cidadão de uma cidade em um grande globo de sorteio. (MAHMUD et al., 2020).

Por sua simplicidade e base teórica sólida, serve como um padrão contra o qual outros métodos mais complexos são frequentemente medidos. Embora o conceito seja simples, implementar uma SRS verdadeira em um conjunto de dados

que abrange terabytes ou petabytes e é distribuído por centenas de servidores é um desafio de engenharia. A escolha de usar a SRS ou suas variantes envolve um trade-off entre pureza teórica e desempenho prático. (DJOUZI; BEGHDAD-BEY; AMAMRA, 2023).

A principal e duradoura força da SRS é sua simplicidade teórica e sua garantia de produzir uma amostra imparcial, desde que seja implementada corretamente. Isso a torna uma linha de base para pesquisas e um método confiável quando a população é relativamente homogênea.

No entanto, a implementação ingênua da SRS requer acesso aleatório a todo o conjunto de dados, o que é impossível em um contexto de streaming e altamente ineficiente em um contexto distribuído. Mais fundamentalmente, como Mahmud et al. (2020) apontam, a SRS tradicional é uma operação em nível de registro, logo, em sistemas de arquivos distribuídos, isso significa acessar potencialmente milhares de blocos de dados diferentes espalhados por um cluster apenas para recuperar alguns registros de cada um, resultando em um gargalo de entrada/saída de rede. Além disso, uma fraqueza estatística significativa é que o SRS, por sua própria natureza, pode facilmente sub-representar ou ignorar completamente subgrupos pequenos, mas extremamente importantes, dentro de um conjunto de dados diversificado.

4.1.2 Reservoir Sampling

Esta é uma família de algoritmos particularmente engenhosa, projetada para situações em que os dados chegam em um fluxo contínuo e seu tamanho total, frequentemente denotado como N, é desconhecido antecipadamente. Imagine a tarefa de selecionar 100 pessoas aleatórias de um desfile de comprimento desconhecido à medida que ele passa. Usando a amostragem de reservatório, você selecionaria as primeiras 100 pessoas para formar sua amostra inicial, ou "reservatório". Para cada pessoa subsequente que passa, digamos, a j-ésima pessoa, você toma uma decisão probabilística. Com uma probabilidade de 100/j, você decide trocar essa nova pessoa por uma das 100 pessoas atualmente em seu reservatório, escolhidas aleatoriamente. (LIU; ZHANG, 2020).

Este procedimento garante que, a qualquer momento, as 100 pessoas em seu reservatório constituam uma verdadeira amostra aleatória simples de todas as pessoas que passaram até o momento. Isso o torna ideal para Data Profiling de

fontes como sensores de IoT ou tráfego da web em tempo real. (DJOUZI; BEGHDAD-BEY; AMAMRA, 2023).

As vantagens são que ele funciona com uma única passagem sobre os dados e não requer conhecimento prévio do tamanho total do dataset, além de ser ideal para amostragem de fluxos de dados, como streaming. Já as desvantagens são que ele é um algoritmo inerentemente sequencial, o que dificulta sua paralelização em ambientes distribuídos. Por fim, suas principais aplicações se resumem a perfilamento e análise de dados em tempo real a partir de fluxos contínuos, como logs de rede ou dados de mídias sociais.

4.1.3 Scalable Simple Random Sampling (ScaSRS)

Para conjuntos de dados massivos e estáticos, ou seja, sem streaming, o ScaSRS surgiu como uma alternativa altamente eficiente e paralelizável aos métodos ingênuos de SRS. Conforme explicado por Djouzi et al. (2022), a principal inovação do ScaSRS é o uso de limiares probabilísticos, em vez de realizar a operação computacionalmente custosa de atribuir um número aleatório a cada registro e, em seguida, classificar todo o conjunto de dados para escolher os principais candidatos, o ScaSRS pode aceitar ou rejeitar rapidamente uma grande fração dos itens com computação mínima. Apenas um pequeno subconjunto de itens "pendentes" precisa ser totalmente classificado, melhorando drasticamente o desempenho e tornando-o adequado para estruturas de processamento paralelo como o Apache Spark. Assim ele consegue tomar uma decisão rápida sobre a maioria dos registros, reduzindo o número de itens que precisam de fato ser ordenados.

O algoritmo calcula limites q1 e q2 e para cada registro do dataset, ele gera uma chave aleatória: se for menor que q2, o registro é aceito; se maior que q1, é rejeitado. Se a chave estiver entre q2 e q1, o registro vai para uma "lista de espera". Após percorrer o dataset, essa lista é ordenada, e os registros restantes para completar a amostra de tamanho k são selecionados dela. Essa abordagem "aceitar/rejeitar/esperar" torna o ScaSRS eficiente para grandes volumes de dados (DJOUZI; BEGHDAD-BEY; AMAMRA, 2022).

O algoritmo é mais rápido e escalável que métodos de SRS baseados em ordenação, ideal para grandes datasets. Ele é projetado para ambientes de computação paralela, oferecendo alto desempenho, o que é essencial para Big

Data. No entanto, ele é mais complexo de implementar que um SRS simples devido ao cálculo de limiares probabilísticos e gestão da lista de espera. Requer conhecimento prévio do tamanho total da população (n) e da amostra desejada (k), o que limita seu uso em cenários de streaming, onde o tamanho do dataset é desconhecido.

Sua principal aplicação é a seleção eficiente de amostras aleatórias simples de datasets massivos. Djouzi et al. (2022) demonstram seu uso em amostragem adaptativa para classificação em Big Data. Ao resolver o desafio do Volume, o algoritmo também impacta indiretamente outros aspectos do Big Data, como velocidade e variedade.

4.1.4 Stratified Sampling

A amostragem aleatória simples trata toda a população como uma entidade única e uniforme. No entanto, a maioria dos grandes conjuntos de dados é heterogênea, contendo subgrupos distintos e significativos. Por exemplo, um conjunto de dados de clientes pode incluir usuários de diferentes países, com diferentes níveis de assinatura e diferentes padrões de atividade. Uma amostra aleatória simples pode, por acaso, super-representar um grupo e mal capturar outro. A amostragem estratificada é uma técnica projetada para evitar isso. Ela envolve um processo de duas etapas: primeiro, toda a população é particionada em subgrupos distintos e não sobrepostos, conhecidos como "estratos". Em seguida, uma amostra aleatória simples é extraída de cada estrato. Isso garante que cada subgrupo definido seja proporcionalmente (ou intencionalmente) representado na amostra final, levando a um quadro geral mais equilibrado e preciso. (ZAYED; BA-ALWI, 2025).

A pesquisa de Anupama e Lakshmi (2022) afirma explicitamente que "a amostragem estratificada demonstrou ter um desempenho superior ao de um método de amostragem aleatória simples" em contextos de big data. Sua principal aplicação é garantir que diferentes segmentos de dados sejam adequadamente representados na amostra, o que, por sua vez, aumenta a precisão de qualquer perfil de dados ou tarefa analítica realizada sobre ela. Por exemplo, para cenários de Big Data e qualidade de dados, a amostragem estratificada pode garantir que o perfilamento inclua dados de todas as fontes ou períodos, evitando que um lote localizado de dados limpos mascare problemas sistêmicos em outros lugares.

A principal vantagem da amostragem estratificada é a melhoria na representatividade e precisão que ela oferece, especialmente para conjuntos de dados com alta variabilidade entre subgrupos. Como diz uma análise, o método "obtém melhor cobertura ao explorar informações adicionais" sobre a estrutura da população, isso leva a estimativas mais confiáveis e perfis de dados mais robustos.

A desvantagem mais significativa é que ele requer conhecimento prévio da população para definir os estratos de forma eficaz. O analista deve saber quais características são importantes para a estratificação antes de começar. Além disso, o próprio processo de estratificação pode ser uma "operação dispendiosa" do ponto de vista computacional, exigindo uma revisão completa dos dados para categorizar cada registro em seu respectivo estrato. Como observam Djouzi et al. (2023), o método "não é útil quando a população não pode ser dividida em subgrupos disjuntos".

4.1.5 Systematic Sampling

Esta é uma técnica simples, na qual um ponto de partida aleatório é escolhido de uma lista e, em seguida, cada k-ésimo elemento é selecionado posteriormente. Por exemplo, um analista pode decidir fazer o perfilamento de um arquivo de log grande após selecionar a primeira entrada e, em seguida, cada 10.000ª entrada subsequente. (DJOUZI; BEGHDAD-BEY; AMAMRA, 2022).

Esse método encontra seu nicho em cenários específicos de big data, por exemplo, ela é particularmente útil para dados ordenados, como logs de séries temporais ou registros transacionais, onde fornece uma maneira fácil de obter uma amostra distribuída uniformemente por toda a linha do tempo.

Como dito por Djouzi (2022) e Zayed (2025), esse é um método geralmente mais simples de implementar do que a amostragem aleatória simples (SRS). A amostragem sistemática tem a vantagem de distribuir "a amostra de forma mais uniforme pela população", o que às vezes pode levar a uma precisão maior do que a SRS. No entanto, a amostragem sistemática é vulnerável a viés se os dados contiverem um padrão periódico oculto que coincide com o intervalo de amostragem, k. Por exemplo, se um sistema gera um relatório de resumo a cada 1000 entradas e o intervalo de amostragem também é 1000, a amostra consistiria inteiramente de relatórios de resumo, criando um resultado completamente não representativo.

4.1.6 Cluster Sampling

Nesse método, a população já está naturalmente agrupada em "clusters", como usuários em diferentes regiões geográficas ou dados armazenados em diferentes servidores. Em vez de amostrar registros individuais de toda a população, o analista seleciona aleatoriamente vários clusters inteiros e, em seguida, analisa cada registro dentro desses clusters escolhidos. (ZAYED; BA-ALWI, 2025).

Esse método também encontra seu nicho em cenários específicos de big data. A amostragem por clusters é altamente relevante para o Data Profiling que são particionados física ou logicamente em um sistema distribuído. Ela pode reduzir drasticamente o "custo" de acesso aos dados, pois requer apenas a leitura de dados de alguns locais (os clusters selecionados), em vez de dados de toda a rede.

Esse é um método mais econômico abordado por Zayed (2025) e Djouzi et al. (2023) do que a amostragem simples aleatória (SRS). Sua principal vantagem está na sua eficiência logística para dados dispersos. No entanto, a principal fraqueza da amostragem por clusters é que ela pode levar a um erro amostral maior se os próprios clusters não forem internamente heterogêneos e representativos da população geral. Se cada cluster for muito homogêneo, mas diferente de outros clusters, a qualidade da amostra dependerá fortemente de quais clusters foram escolhidos aleatoriamente.

4.1.7 Bernoulli Sampling

Um outro destaque é para a amostragem de Bernoulli, uma técnica de amostragem probabilística em nível de registro onde cada item do dataset tem a mesma probabilidade “p” de ser incluído na amostra. O processo funciona percorrendo todos os registros da população, onde para cada registro, uma decisão independente é tomada (como lançar uma moeda com probabilidade “p” de dar "cara") para determinar se aquele registro fará parte da amostra. Diferente da Amostragem Aleatória Simples (SRS) que seleciona um número fixo “k” de itens, na amostragem de Bernoulli, o tamanho final da amostra não é pré-determinado. (MAHMUD et al., 2020).

Suas vantagens estão no fato de que a seleção de cada item é um evento independente, o que simplifica a análise teórica do processo de amostragem. No entanto, a principal desvantagem é que o tamanho final da amostra é aleatório e não

fixo. Isso torna difícil estimar a latência de processamento e planejar a alocação de recursos computacionais. E por não ter um tamanho fixo, pode ser menos prático para certas aplicações que dependem de um número exato de amostras. A Amostragem Aleatória Simples é frequentemente usada para superar essa limitação.

A amostragem de Bernoulli é crucial em computação aproximada, como no BlinkML, onde gerencia grandes datasets que excedem a memória. Ela também é um modelo teórico vital para analisar mecanismos de amostragem e viés de seleção em Big Data. (KIM; WANG, 2019).

4.2 Amostragem Avançada em nível de Bloco

À medida que o processamento de big data migrou de máquinas únicas e monolíticas para clusters vastos e distribuídos, técnicas de amostragem fundamentais tiveram que evoluir. A própria arquitetura de plataformas de dados modernas, como Apache Hadoop e Spark, exigiu uma mudança de paradigma na abordagem da amostragem, assim, esta seção detalha essa evolução crucial, passando da da amostragem centrada em registros para o poder dos designs arquitetônicos com reconhecimento de blocos.

4.2.1 Random Sample Partition (RSP)

O modelo de Partição de Amostra Aleatória (RSP) é uma solução arquitetônica para o dilema entre eficiência e validade estatística em amostragens distribuídas. Ele representa uma mudança no pensamento, em que o próprio modelo de dados é projetado com análises e perfilamentos em mente.

Conforme detalhado no trabalho de Salloum, Huang e He (2019), o modelo RSP redefine como um grande conjunto de dados é armazenado. Em vez de simplesmente dividir um arquivo em blocos sequenciais, o modelo RSP representa um grande conjunto de dados como "um conjunto de subconjuntos de dados não sobrepostos, chamados blocos de dados RSP, onde cada bloco de dados RSP tem uma distribuição de probabilidade semelhante à de todo o grande conjunto de dados". Esta é uma distinção crucial, pois um RSP é criado por meio de um processo offline único que reconstrói os blocos de dados de forma inteligente. Cada novo bloco RSP é formado pela coleta de pequenas fatias aleatórias de dados de todos os blocos HDFS originais e sua combinação. Esse processo garante que cada

bloco RSP seja, por sua própria construção, uma amostra aleatória simples estatisticamente válida de todo o conjunto de dados. Ou seja, o processo de duas etapas envolve embaralhar os registros dentro dos blocos HDFS originais e, em seguida, fatiar e recombinar essas fatias para formar os novos blocos RSP, que são os blocos prontos para uso já pré-processados de forma aleatória.

O modelo RSP oferece principalmente escalabilidade, permitindo a análise de conjuntos de dados massivos em pequenos clusters de hardware de commodities. Além disso, ele potencializa a eficiência das operações em nível de bloco, ao mesmo tempo em que fornece garantias estatísticas robustas.

A principal limitação é a exigência de um custo computacional inicial e único para criar o RSP a partir dos dados originais. O modelo também foi projetado para processamento em lote de conjuntos de dados estáticos e não é adequado para streaming de dados em tempo real. Por fim, ele apresenta limitações inerentes para tarefas que exigem uma visão global dos dados, como a detecção de registros duplicados que possam existir em diferentes blocos do RSP.

Uma aplicação forte dessa técnica é que ela gerou avanços recentes na criação do método RSP-Explore, uma estrutura prática construída sobre o modelo RSP, projetada para capacitar cientistas de dados a executar tarefas de perfilamento importantes iterativamente, mesmo em clusters de computação pequenos e com recursos limitados. O método operacionaliza os benefícios do RSP por meio de um fluxo de trabalho simples para três funções principais: estimativa estatística, com a análise de alguns blocos RSP em paralelo para estimar rapidamente as estatísticas do conjunto de dados com fortes garantias de precisão; detecção de erros, com a coleta de amostras de blocos RSP para avaliar rapidamente a qualidade dos dados, identificando erros, outliers e valores ausentes; e limpeza de dados, com operações de limpeza a uma pequena amostra de blocos RSP sujos para estimar as propriedades estatísticas de todo o conjunto de dados limpo e desconhecido, evitando o custo total da limpeza.

4.2.2 I-Sampling

Embora o RSP represente uma solução arquitetônica abrangente, outros métodos especializados baseados em blocos foram desenvolvidos para abordar desafios específicos. Proposto por He et al. (2017), o I-sampling é outro método baseado em blocos projetado para conjuntos de dados de grande escala já

armazenados em um sistema distribuído. Ele é descrito como um componente-chave dentro de uma "estrutura de aprendizagem de conjunto assintótico" mais ampla. Esse contexto sugere que o I-sampling foi projetado para facilitar a construção iterativa de modelos, onde amostras são usadas para treinar uma série de modelos cujos resultados são então combinados, além de ser descrito como um método promissor para tarefas de Data Profiling em máquinas distribuídas.

O I-Sampling envolve um processo de várias etapas, o grande conjunto de dados é inicialmente dividido em K blocos de dados não sobrepostos, chamados de "blocos primários" (A1, A2, ..., AK). Os registros dentro de cada bloco primário são embaralhados aleatoriamente e isso cria um novo conjunto de "blocos embaralhados" (B1, B2, ..., BK). O objetivo deste embaralhamento é quebrar qualquer ordem local ou global que existisse nos dados originais, garantindo que os passos seguintes operem sobre dados randomizados internamente. Então, em vez de usar os blocos embaralhados diretamente, o I-Sampling cria um "pool" de novos blocos, chamados de "blocos base" (C1, C2, ..., CL). Cada bloco base é construído de uma forma engenhosa: ele é formado pela seleção aleatória de um pequeno número de registros de cada um dos blocos embaralhados (B). Isso significa que cada bloco base é uma representação do dataset inteiro, contendo pequenas amostras aleatórias de todas as suas partes. Por fim, a amostra final é criada selecionando-se aleatoriamente um ou mais blocos base do pool gerado na etapa anterior. Como cada bloco base já é, por construção, uma amostra representativa do todo, qualquer bloco selecionado do pool também será uma amostra de alta qualidade.

Liu e Zhang (2020) confirmam que os resultados experimentais demonstram que a distribuição de dados produzida pelo I-Sampling é aproximadamente a mesma que a do conjunto de dados original, validando sua eficácia como uma técnica de amostragem representativa para ambientes distribuídos.

Sua principal vantagem é a capacidade de lidar com datasets onde os registros não estão ordenados de forma aleatória. A etapa de embaralhamento foi projetada especificamente para mitigar o viés que surgiria ao amostrar blocos de dados ordenados. Além disso, a teoria por trás do I-Sampling demonstra que os blocos base gerados são estimadores imparciais e consistentes da média e da variância do dataset original. As operações de embaralhamento e geração de blocos podem ser realizadas de forma independente e paralela em diferentes nós do

cluster, o que o torna altamente escalável. Por fim, o resultado do processo é um "pool" de blocos base, cada um sendo uma amostra independente e de alta qualidade. Isso é útil para tarefas que se beneficiam de múltiplas amostras, como validação cruzada ou ensemble learning.

No entanto, o fluxo de trabalho de quatro etapas: particionar, embaralhar, gerar e amostrar é mais complexo do que uma amostragem de passo único. E embora escalável, o processo de embaralhar todos os blocos primários e recombinar registros para formar os blocos base tem um custo computacional e de I/O. É um investimento inicial para garantir a qualidade estatística das amostras finais. Além disso, a configuração do I-Sampling depende de parâmetros como o número de blocos primários (K), o número de blocos base no pool (L) e o "fator de recombinação", que podem precisar de ajuste para otimizar o trade-off entre custo e precisão.

Por gerar amostras estatisticamente representativas, o I-Sampling é ideal e criado justamente para ambientes de Big Data. Além disso, para tarefas de Data Profiling, como estimar estatísticas descritivas ou entender a distribuição de valores de atributos em um grande dataset ele pode ser útil. Além disso, é a técnica de escolha quando se suspeita que os dados estão ordenados por algum critério, como data ou ID de cliente, pois a amostragem de blocos simples falharia em capturar a diversidade do dataset. Por fim, o I-Sampling pode gerar datasets de treinamento para algoritmos de machine learning que resultam em um desempenho de modelo praticamente idêntico ao que seria obtido com uma amostragem aleatória simples (SRS), mas de uma forma que é viável e escalável para Big Data. (HE et al., 2017)

4.2.3 CDFRS

CDFRS é uma abordagem de amostragem escalável que gera amostras com garantia de preservação da distribuição, medida pela distância de Kolmogorov-Smirnov entre as Funções de Distribuição Acumulada - CDF. A principal inovação do CDFRS é que ele não apenas seleciona uma amostra, mas o faz com uma garantia teórica de que a distribuição da amostra será próxima da distribuição do dataset original. Essa "proximidade" é formalmente definida: uma amostra é considerada uma "Amostra Aleatória CDF" se a distância máxima entre a sua Função de Distribuição Acumulada (CDF) e a CDF do dataset completo for menor que um limite de erro pré-definido. Para alcançar isso de forma eficiente em Big

Data, o CDFRS utiliza uma abordagem inteligente de amostragem em múltiplos estágios, todos operando em nível de bloco. (CAI et al., 2024).

Na primeira etapa o algoritmo começa com todos os blocos de dados do dataset no HDFS. Ele então particiona aleatoriamente esses blocos em subconjuntos de tamanho igual, ou seja, é a primeira camada de aleatorização. Na segunda etapa o algoritmo seleciona aleatoriamente um desses subconjuntos de blocos e este subconjunto, chamado D, é uma amostra em nível de bloco do dataset inteiro. O tamanho deste subconjunto (K blocos) é calculado para garantir que, com alta probabilidade, sua distribuição já seja próxima à do dataset original. No entanto, este subconjunto ainda é grande e seus blocos não têm garantia de aleatoriedade interna. Por fim, na terceira etapa, considerada a mais importante, ao invés de aplicar o caro algoritmo RSP no dataset inteiro, o CDFRS o aplica apenas no subconjunto de blocos D selecionado anteriormente. O RSP reorganiza os registros dentro desses K blocos, gerando K novos blocos de saída (B'k), onde cada um é uma amostra aleatória e representativa dos dados contidos em D.

Conforme abordado por CAI et al. (2024), é justamente essa abordagem de dois níveis (uma amostragem de bloco "grosseira" para reduzir o escopo, seguida por uma amostragem de bloco "refinada" com RSP no subconjunto) que permite ao CDFRS ser extremamente rápido e, ao mesmo tempo, estatisticamente robusto.

A maior vantagem do CDFRS é que ele é um dos poucos métodos que oferece uma prova teórica de que a amostra gerada preserva as características de distribuição do dataset original dentro de um limite de erro controlável. Além disso, os experimentos feitos por CAI et al. (2024) mostram que o CDFRS supera outras técnicas de amostragem distribuídas em ordens de magnitude, pois ele consegue amostrar um dataset de 10TB em centenas de segundos, enquanto métodos concorrentes levam dezenas de milhares de segundos ou falham por falta de memória. Outra vantagem é que assim como o RSP, o CDFRS pode ser aplicado a datasets distribuídos com qualquer layout inicial de registros, ordenado ou não, pois seus múltiplos estágios de aleatorização garantem uma amostra final randomizada.

Suas principais desvantagens estão no fato de que é uma técnica multi-etapas que combina partição de blocos, seleção e o algoritmo RSP, logo, sua implementação é mais complexa do que métodos mais simples, além da qualidade da amostra final ser dependente da escolha dos parâmetros iniciais de limite de erro e probabilidade de confiança, que determinam o número de blocos (K) a serem

selecionados na segunda etapa. Logo, uma escolha inadequada pode afetar o equilíbrio entre velocidade e precisão.

O CDFRS é mais indicado para os cenários desafiadores de Big Data, como análise de datasets na escala de terabytes e petabytes, onde a eficiência computacional é o fator mais crítico e outras técnicas falham ou se tornam impraticáveis. Pode ser usado em machine learning em larga escala, em que modelos treinados com pequenas amostras do CDFRS alcançam precisão quase idêntica a modelos treinados no dataset completo, mas em uma fração mínima do tempo. Por fim, também é usado para determinação eficiente do tamanho da amostra, em que o próprio autor propõe o algoritmo A², que utiliza as amostras geradas pelo CDFRS para determinar o tamanho de amostra ideal para uma tarefa, de forma mais eficiente do que métodos tradicionais que exigem o treinamento de múltiplos modelos.

4.3 Outras Amostragens

Visando explorar as técnicas de amostragem em Big Data, este trabalho se aprofunda nas seções 4.1 e 4.2, que abordam métodos comuns e amplamente empregados. Contudo, reconhecemos a existência de outros grupos menos conhecidos, mas igualmente relevantes. Devido à limitação de tempo, faremos uma breve menção a esses grupos, descrevendo suas características e potenciais benefícios, o que poderá servir de base para futuras análises detalhadas. Assim, serão abordadas como um grande grupo, mencionando as técnicas que os compõem e sua aplicabilidade em Big Data.

Além das inovações arquitetônicas que otimizam a amostragem para sistemas distribuídos, outra fronteira de pesquisa concentra-se em tornar o próprio processo de amostragem mais "inteligente". Essa inteligência se manifesta de duas maneiras principais: primeiro, em algoritmos que se adaptam dinamicamente aos dados para alcançar eficiência computacional; segundo, em métodos projetados para otimizar a eficiência cognitiva do cientista de dados humano. Esta seção explora essas técnicas avançadas, que se afastam de regras de amostragem estáticas e buscam processos de seleção dinâmicos, orientados por dados e metas.

Essa evolução revela uma bifurcação fascinante na filosofia do que significa "amostragem inteligente". De um lado, há uma abordagem centrada na máquina,

focada na otimização estatística e computacional, cujo objetivo é fazer com que o algoritmo execute seu trabalho com mais eficiência. De outro, há uma abordagem centrada no ser humano, focada na otimização cognitiva, cujo objetivo é ajudar o analista a gerar insights com mais eficácia.

4.3.1 Adaptive Sampling

Os métodos tradicionais de amostragem exigem que o usuário especifique o tamanho de amostra desejado com antecedência. Isso geralmente é um tiro no escuro, podendo levar a uma amostra muito pequena para ser precisa ou excessivamente grande. A amostragem adaptativa resolve esse problema com elegância. O princípio básico é começar com uma pequena amostra inicial e, em seguida, expandi-la iterativamente, passo a passo. A decisão de parar de adicionar mais dados não é arbitrária; ela se baseia no monitoramento de uma variável de interesse, normalmente o desempenho de um modelo de aprendizado de máquina, como sua precisão de classificação. O processo continua até que essa métrica se estabilize ou "convirja", indicando que a adição de mais dados provavelmente não melhora o resultado significativamente. Isso garante que a amostra final seja grande o suficiente para ser eficaz, proporcionando um equilíbrio ideal entre precisão e custo computacional. (DJOUZI; BEGHDAD-BEY; AMAMRA, 2023)

A evolução desses algoritmos, conforme documentado por Djouzi et al. (2022, 2023), mostra um claro avanço em sofisticação e eficiência. As primeiras abordagens, como a amostragem aritmética e geométrica, aumentavam a amostra com incrementos fixos ou proporcionais. Um salto significativo ocorreu com a introdução de algoritmos dinâmicos. O DASA (Dynamic Adaptive Sampling Algorithm) foi pioneiro ao usar a desigualdade de Chernoff para estimar de forma inteligente o número de instâncias adicionais necessárias. Para refinar essa estimativa, o GDAS (Generalized Dynamic Adaptive Sampling) substituiu a abordagem do DASA pela amostragem bootstrap e pela desigualdade de Chebyshev, corrigindo a tendência conservadora do método anterior e resultando em amostras menores. A evolução mais recente, o SDBGDAS, otimizou o GDAS para ambientes de Big Data, incorporando um mecanismo de seleção paralelizável (ScaSRS) para garantir escalabilidade e velocidade em grandes volumes de dados.

O principal ponto forte dos métodos adaptativos é sua capacidade de determinar automaticamente um tamanho de amostra apropriado, o que proporciona

um excelente equilíbrio entre precisão analítica e custo de processamento. Os algoritmos mais recentes, como o SDBGDAS, demonstram ser altamente escaláveis, rápidos e robustos, superando métodos anteriores tanto em tempo quanto no número de instâncias necessárias para convergir.

No entanto, esses algoritmos são inerentemente mais complexos de implementar do que os métodos de amostragem estática. Versões anteriores, que dependiam de técnicas computacionalmente mais intensivas para estimativa de variância, podiam sofrer com sobrecarga significativa em tempo de execução, um problema para o qual métodos como o SDBGDAS foram projetados especificamente.

4.3.2 Active Learning Sampling

A segunda forma de amostragem inteligente muda completamente o objetivo. Como demonstra de forma convincente a pesquisa de Rojas et al. (2017), o objetivo da amostragem nem sempre é criar uma representação perfeitamente representativa do conjunto de dados completo. No contexto da análise exploratória de dados (EDA), o objetivo principal é ajudar um analista humano a gerar insights. Para tanto, diferentes técnicas de amostragem podem ser utilizadas para enviesar intencionalmente a amostra de maneiras que destaquem aspectos diferentes e interessantes dos dados, otimizando assim o tempo e os recursos cognitivos limitados do analista.

O estudo de Rojas et al. (2017) investigou diversas técnicas desse tipo, emprestadas do campo da aprendizagem ativa. A amostragem por densidade, por exemplo, foca nas tendências gerais ao priorizar pontos em regiões densas e comuns do conjunto de dados. Já a amostragem por incerteza muda o foco para casos ambíguos, selecionando os pontos que um modelo classificador tem maior dificuldade em rotular, como outliers ou instâncias em fronteiras de decisão. Uma evolução desta última é a consulta por comitê (QBC), que utiliza múltiplos classificadores e seleciona os pontos onde eles mais discordam, sendo eficaz em revelar áreas de alta complexidade inerentes aos dados.

Esses métodos capacitam os cientistas de dados, fornecendo-lhes um conjunto de ferramentas para direcionar sua atenção limitada a características específicas dos dados, sejam tendências gerais, valores discrepantes incomuns ou casos de contorno complexos. A pesquisa de Rojas et al. (2017) constatou que

fornecer aos analistas acesso a múltiplas técnicas de amostragem pode reduzir o tempo de exploração de dados e melhorar a qualidade geral e a amplitude de seus insights. A limitação fundamental, no entanto, é crucial: o viés é uma característica deliberada, não um erro. Essas amostras são projetadas para a exploração e a geração de hipóteses, sendo inerentemente inadequadas para realizar inferências estatísticas sobre a população total.

4.3.3 Bias Correction Sampling

Uma terceira dimensão crítica da inteligência na amostragem aborda um profundo desafio inerente à maioria das fontes de big data. Esses dados são "dados encontrados" ou "dados de oportunidade", registros da web, postagens em mídias sociais ou registros de transações que não foram gerados por meio de um experimento aleatório cuidadosamente controlado. Isso significa que os dados estão quase sempre sujeitos a "viés de seleção".

Por exemplo, avaliações de produtos são deixadas apenas por um determinado tipo de cliente, e os dados de mídias sociais refletem apenas as opiniões daqueles que usam uma plataforma específica. Como Kim e Wang (2019) enfatizam, é "fundamental ajustar o viés de seleção" nesses dados se se deseja tirar conclusões válidas e generalizáveis sobre uma população mais ampla.

O trabalho de Kim e Wang (2019) propõe técnicas estatísticas sofisticadas para corrigir esse viés inerente, como a amostragem inversa, método que trata a amostra enviesada de big data como um dado e busca corrigir suas falhas por meio de uma subamostragem inteligente a partir dela. A chave é o uso de informações auxiliares de uma fonte externa e confiável, como dados censitários. Assim, esses dados externos fornecem totais populacionais conhecidos para determinadas variáveis, por exemplo, distribuição por idade e gênero. Esses totais conhecidos são usados para calcular pesos de importância para cada registro na amostra do big data. Registros de dados demográficos super-representados no big data recebem pesos baixos, enquanto aqueles de dados demográficos sub-representados recebem pesos altos. Uma amostra final corrigida é então extraída do big data com uma probabilidade de seleção proporcional a esses pesos, rebalanceando efetivamente a amostra para corresponder à população real. Ou seja, na primeira fase, uma amostra é obtida do Big Data; enquanto na segunda fase, uma subamostra é selecionada da primeira, com probabilidades de seleção inversamente

proporcionais a alguma medida de "importância" ou viés, calibrada usando uma fonte de dados externa e confiável.

A vantagem desses métodos é justamente permitir corrigir o viés de seleção que pode estar presente em conjuntos de Big Data. No entanto, sua desvantagem está ligada ao fato de que requer uma fonte de dados auxiliar (externa e probabilística) para a calibração, o que nem sempre está disponível. De modo geral sua aplicação é voltada para quando se tem um grande volume de dados de uma fonte não confiável (ex: mídias sociais) e uma pequena amostra de pesquisa de alta qualidade. A amostra de pesquisa pode ser usada para calibrar e corrigir os resultados da amostra maior. Vemos como o trabalho de Kim e Wang (2019) é dedicado a corrigir o "viés de seleção", reconhecendo que muitas fontes de big data são inerentemente amostras não representativas e não probabilísticas para começar.

4.3.4 Imbalanced Data Sampling

Um cenário comum e desafiador em perfis de dados e aprendizado de máquina é o desequilíbrio de classes, em que uma classe de interesse é amplamente sub-representada no conjunto de dados. Técnicas de amostragem padrão frequentemente produzem uma amostra em que a classe minoritária ainda é rara ou mesmo totalmente ausente, logo, técnicas de amostragem especializadas são necessárias para lidar com isso.

Zhu et al. (2024) propõem uma abordagem híbrida sofisticada que combina a Amostragem Mista por Jogo com um algoritmo genético. O cerne desse método é o uso de uma "ideia de jogo" para determinar dinamicamente a combinação ideal de sobreamostragem (criando mais instâncias da classe minoritária) e subamostragem (removendo instâncias da classe majoritária) para o conjunto de dados específico em questão.

Sua vantagem está no fato de que ela adapta-se às características do dataset para encontrar a melhor estratégia de amostragem, além de melhorar o equilíbrio do dataset e aumentar a diversidade das características dos dados. No entanto, sua desvantagem se dá por ser um processo computacionalmente complexo, pois envolve testar múltiplas combinações de métodos e proporções. Por fim, sua principal aplicação é no contexto de perfilamento e classificação de dados

desbalanceados, como por exemplo a detecção de fraude de cartão de crédito ou detecção de doenças raras em uma classe alvo.

4.3.5 Non Probability based Sampling

Métodos de amostragem não probabilística são abordagens nas quais indivíduos são selecionados da população sem o princípio da randomização, o que significa que cada membro não tem uma chance conhecida ou igual de ser incluído. Esses métodos são frequentemente empregados quando a amostragem probabilística é impraticável devido a restrições como recursos limitados, dificuldades de acesso ou a natureza da pesquisa exploratória, onde a generalização não é o foco principal. (ZAYED; BA-ALWI, 2025).

Embora os métodos de amostragem não probabilística sejam menos robustos em termos de produção de resultados estatisticamente generalizáveis, eles podem ser valiosos para gerar insights iniciais, compreender grupos específicos ou coletar dados em contextos onde a aleatoriedade não é viável.

A amostragem por conveniência é um um método que seleciona os participantes com base na facilidade de acesso e disponibilidade para o pesquisador, também sendo conhecida como amostragem acidental. Já a amostragem por quotas é um método onde a população é dividida em subgrupos exclusivos e um número pré-definido de participantes é selecionado de cada grupo para garantir a representação de características específicas. A amostragem por julgamento é uma técnica onde o pesquisador seleciona os participantes deliberadamente, com base em seu próprio julgamento e em critérios específicos relevantes para a questão da pesquisa. Já a amostragem por bola de neve é um método de recrutamento por indicação, onde os participantes iniciais ajudam a encontrar outros participantes adequados para o estudo. (ZAYED; BA-ALWI, 2025).

De modo geral as principais vantagens são a velocidade e o baixo custo. Porém, a desvantagem avassaladora é o alto risco de viés de amostragem, tornando impossível generalizar os resultados para a população em geral com confiança estatística. E embora úteis para pesquisa exploratória ou geração de hipóteses, esses métodos são explicitamente observados como desinteressantes devido à sua subjetividade e baixa automação, tornando-os inadequados para mineração de dados e aplicações de big data em cenários muito complexos. Esta perspectiva crítica é essencial para um guia de iniciantes para evitar o uso indevido.

5. Discussão

A seleção de uma técnica de amostragem apropriada para usar em ambientes Big Data não é uma decisão única. Envolve navegar por uma complexa rede de compensações que equilibram restrições computacionais e objetivos estatísticos e analíticos. As diversas técnicas discutidas nesta revisão ocupam, cada uma, uma posição única nesse cenário, oferecendo vantagens distintas ao custo de limitações específicas. Uma comparação sistemática dessas compensações é essencial para qualquer profissional que busque fazer uma escolha informada.

Conforme mencionado, este trabalho não busca impor ou favorecer uma técnica em detrimento de outra, sendo levantado as definições, vantagens e recomendação de suas aplicações, logo, a decisão de utilizá-las em contextos de Big Data ou em conjunto com tarefas de Data Profiling cabe exclusivamente ao cientista de dados após um entendimento do cenário.

A escolha do método ideal não é universal, mas um trade-off ditado pelos objetivos analíticos e pelas restrições de infraestrutura. Este estudo identificou uma clara trajetória evolutiva, impulsionada pelas plataformas de Big Data. A impraticabilidade da amostragem em nível de registro em sistemas distribuídos levou à adoção de métodos baseados em blocos. O risco de viés inerente a essa abordagem, por sua vez, estimulou a criação de técnicas mais sofisticadas e com consciência de arquitetura, como o RSP, o I-Sampling e o CDFRS, que garantem representatividade em ambientes distribuídos. Isso demonstra que as técnicas modernas não podem ser dissociadas da arquitetura em que operam.

Observa-se também uma tensão fundamental no propósito da amostragem: de um lado, a busca por uma representação do todo; de outro, a exploração direcionada para descobrir padrões específicos. Essa dualidade exige que o cientista de dados domine um portfólio de técnicas, aplicando métodos como a Amostragem Estratificada, a Amostragem Mista de Jogos ou a Amostragem Inversa para garantir a representação de classes desbalanceadas ou a Aprendizagem Ativa para focar em pontos de dados mais informativos, como anomalias. O surgimento de métodos adaptativos representa o amadurecimento da área, mudando o foco da questão estática de "como amostrar" para a otimização dinâmica de "quando parar de amostrar", equilibrando custo e utilidade analítica.

Ademais, as fronteiras deste campo continuam a se expandir, com aplicações emergentes como o uso do I-Sampling na criação de conjuntos de treino para LLMs

e a combinação de múltiplas abordagens, como visto no CDFRS. Em suma, a amostragem se firma como uma disciplina dinâmica e essencial, onde o valor extraído do Big Data depende cada vez menos do volume processado e mais da inteligência da estratégia de amostragem empregada.

5.1 Relação com Data Profiling

Apesar de cada técnica ter sido abordada individualmente, e de suas aplicações em Big Data e as possibilidades em Machine Learning, como o I-Sampling, terem sido compreendidas, o Data Profiling ainda carece de detalhamento. Logo, para sanar uma potencial lacuna entre a teoria da amostragem e sua aplicação prática, vamos conectar resumidamente as técnicas de amostragem discutidas com as tarefas de Data Profiling detalhadas no Capítulo 2. A escolha de uma técnica de amostragem não deve ser arbitrária, mas sim uma decisão estratégica que considera o objetivo do perfilamento, seja ele em uma única coluna, em múltiplas colunas ou na descoberta de dependências complexas.

Para as tarefas de perfilamento de coluna única, como a análise de distribuição de valores, cálculo de cardinalidade ou identificação de valores nulos, a principal exigência é a representatividade estatística. Técnicas como Simple Random Sampling (SRS) e suas variantes escaláveis como o ScaSRS são um ponto de partida eficaz, contanto que o dataset não possua vieses estruturais. No entanto, em ambientes Big Data distribuídos, abordagens com reconhecimento de blocos como RSP, I-Sampling e, principalmente, CDFRS são superiores. O CDFRS, por exemplo, oferece a garantia teórica de que a distribuição da amostra é fiel à do dataset original, tornando-o ideal para estimar com precisão a frequência de valores e a porcentagem de nulos sem analisar todos os dados. Em cenários de dados ordenados, como logs de séries temporais, o Systematic Sampling garante uma cobertura uniforme, evitando que o perfilamento se concentre em um único período.

Ao avançar para o perfilamento de múltiplas colunas, que busca descobrir correlações, clusters e outliers, a escolha da amostragem se torna mais crítica. Uma amostra aleatória simples pode facilmente ignorar outliers ou subgrupos raros, mas importantes. Aqui, o Stratified Sampling é fundamental se houver conhecimento prévio de segmentos importantes nos dados (ex: tipos de clientes, regiões geográficas), garantindo que as correlações sejam analisadas dentro de cada estrato relevante. Para a descoberta exploratória de outliers e padrões incomuns, as

técnicas de Active Learning Sampling são as mais indicadas. A amostragem por incerteza, por exemplo, foca deliberadamente em pontos de dados que são difíceis de classificar, que são frequentemente os outliers ou registros em fronteiras de decisão, otimizando o tempo do analista para a descoberta de anomalias.

Finalmente, para a descoberta de dependências, como chaves candidatas (unicidade) e chaves estrangeiras (dependências de inclusão), a amostragem apresenta seu maior desafio, pois essas são propriedades globais do dataset. É inviável validar uma chave primária com total certeza em uma amostra. Contudo, a amostragem pode atuar como uma poderosa ferramenta de geração de hipóteses. Utilizando uma amostra grande e de alta representatividade, gerada por métodos como CDFRS ou RSP, um analista pode executar algoritmos de descoberta de chaves. As chaves candidatas identificadas na amostra podem, então, ser validadas de forma muito mais eficiente no dataset completo, em vez de executar a custosa busca inicial em terabytes de dados. Para dados com problemas conhecidos, como classes desbalanceadas ou viés de seleção, técnicas como Imbalanced Data Sampling e Bias Correction Sampling devem ser aplicadas antes do perfilamento para garantir que as métricas calculadas (sejam de coluna única ou múltiplas) reflitam a realidade corrigida e não a distribuição enviesada dos dados.

A Tabela 7, apresentada a seguir, sintetiza as informações discutidas, exibindo as respostas às questões de pesquisa RQ1-RQ3. Além disso, traz de forma resumida uma orientação sobre qual técnica é mais indicada para cada tipo de cenário, bem como as recomendações de seu uso em cenários de Data Profiling.

Tabela 7 – Resumo das técnicas de amostragens

Técnica	Princípio	Vantagens	Desvantagens	Aplicação	Data Profiling
Simple Random Sampling (SRS)	Cada item na população tem uma probabilidade igual e conhecida de ser selecionado.	Simplicidade fundamental.	Ineficiente para big data; requer acesso aleatório aos dados e conhecimento do tamanho da amostra.	Linha de base para comparação; datasets de pequeno a médio porte em um único computador.	Perfilamento de coluna única (estatísticas básicas, valores nulos) em datasets homogêneos e de tamanho gerenciável.
Scalable Simple Random Sampling (ScaSRS)	SRS otimizado com limiares probabilísticos para reduzir a etapa de ordenação dos dados.	Desempenho e escalabilidade; paralelização; confiabilidade	Implementação mais complexa; requisito de conhecimento prévio do tamanho total do dataset e amostra	Amostragem de datasets massivos onde o SRS tradicional é computacionalmente inviável.	Perfilamento de coluna única em larga escala, quando a velocidade na seleção da amostra de um dataset massivo é crucial.
Stratified Sampling	Divide a população em subgrupos (estratos) e realiza	Garante representatividade de todos os	Exige conhecimento prévio da população para estratificar ;	Populações heterogêneas onde é importante garantir	Perfilamento de coluna única e múltiplas colunas em

	amostragem aleatória dentro de cada um.	subgrupos ; maior precisão com amostras menores.	processo de estratificação pode ser complexo e caro.	a representação de diferentes segmentos de dados.	dados heterogêneos para garantir a análise precisa de todos os segmentos, incluindo a proporção correta de valores nulos e outliers por estrato.
Reservoir Sampling	Seleciona uma amostra de tamanho fixo k de um fluxo de dados de tamanho desconhecido ou muito grande, em uma única passagem	Ideal para fluxos de dados (streams) e datasets de tamanho desconhecido; baixo uso de memória.	Algoritmo sequencial, difícil de paralelizar ; pode ser lento para grandes volumes de dados.	Mineração de dados em grande escala e análise de fluxos de dados (data streams).	Perfilamento de coluna única em tempo real sobre fluxos de dados (data streams), ideal para monitorar estatísticas como média, mínimo/máximo de forma contínua.
Systematic Sampling	Seleciona itens em intervalos regulares (a cada k-ésimo item) a partir de um ponto inicial aleatório.	Simples de implementar; distribui a amostra uniformemente pela população.	Suscetível a viés se houver um padrão periódico nos dados que coincida com o intervalo de amostragem.	Útil quando os dados estão ordenados e não possuem periodicidade oculta.	Perfilamento de coluna única em dados com ordenação natural (ex: séries temporais), garantindo que a análise cubra todo o escopo temporal ou sequencial.
Cluster Sampling	Divide a população em grupos (clusters) e seleciona aleatoriamente clusters inteiros para a amostra.	Reduz custos e complexidade logística em populações grandes e dispersas.	Maior erro amostral se os clusters forem muito homogêneos internamente; os clusters podem não representar bem a população.	Populações grandes e geograficamente dispersas, onde a amostragem individual é impraticável.	Perfilamento de alto nível para estimativas gerais (ex: média de valores em todo o dataset) em sistemas fisicamente distribuídos, reduzindo o custo de acesso aos dados.
Bernoulli Sampling	Cada item tem uma probabilidade igual e independente de ser incluído na amostra.	Simplicidade na seleção, onde cada item é avaliado de forma independente.	O tamanho final da amostra é aleatório e não fixo, o que dificulta o planejamento de recursos e tempo.	Cenários de big data onde o tamanho exato da amostra não é um requisito estrito.	Perfilamento de coluna única, especialmente em cenários de computação aproximada, onde um tamanho de amostra fixo não é um pré-requisito.
Random Sample Partition (RSP)	Particiona o dataset em blocos, onde cada bloco é uma amostra aleatória de todo o conjunto.	Reduz o tempo de amostragem de horas para segundos; evita viés da amostragem por bloco.	A criação inicial dos blocos (particionamento) é uma operação offline que pode ser demorada.	Análise exploratória e limpeza de big data em clusters, permitindo o uso de algoritmos sequenciais em paralelo sobre os blocos.	Perfilamento de coluna única e múltiplas colunas em ambientes distribuídos. Excelente para análise exploratória iterativa, estimativa de distribuições e descoberta de correlações.
I-Sampling	Método baseado em blocos que embaralha registros internamente antes de gerar um "pool de blocos" para a amostragem final.	Funciona bem para datasets com registros ordenados ; boa extensibilidade em sistemas distribuídos.	Assume que os blocos primários têm tamanhos iguais, o que pode exigir ajustes.	Datasets em larga escala onde a ordem dos registros pode enviesar a amostragem baseada em blocos tradicionais.	Perfilamento de coluna única e múltiplas colunas, sendo particularmente útil para datasets distribuídos onde os dados podem estar ordenados, mitigando

					o viés de bloco.
CDFRS	Abordagem escalável que combina amostragem em nível de bloco com o algoritmo RSP em um subconjunto de blocos para garantir a preservação da distribuição.	Extremamente rápido para datasets na escala de terabytes ; supera outras abordagens distribuídas em ordens de magnitude.	A qualidade da amostra depende do número de blocos selecionados na primeira fase, que por sua vez depende dos limites de erro e confiança desejados.	Amostragem ultrarrápida (terabytes em segundos) em sistemas distribuídos (Spark), preservando a distribuição.	Perfilamento avançado de coluna única (distribuição de valores, cardinalidade) e múltiplas colunas (correlações) em datasets de terabytes, com garantia de preservação da distribuição original.
Adaptive Sampling	O processo de seleção da amostra se ajusta com base nas observações feitas durante a própria amostragem.	Determina automaticamente o tamanho ideal da amostra, proporcionando um bom equilíbrio entre precisão analítica e custo computacional.	Algoritmicamente mais complexo que métodos estáticos; alguns métodos (GDAS) podem ser computacionalmente caros.	Tarefas de classificação em big data, otimizando o tamanho da amostra para atingir a convergência do modelo.	Perfilamento iterativo para determinar o tamanho de amostra mínimo necessário para obter uma estimativa estável de uma métrica específica (ex: porcentagem de valores nulos).
Active Learning Sampling	Seleciona os pontos de dados mais informativos (geralmente aqueles com maior incerteza para o modelo) para a rotulagem.	Concentra a atenção limitada de um cientista de dados em recursos de dados específicos (por exemplo, tendências gerais vs. outliers vs. casos complexos).	Produz amostras tendenciosas por design; não é adequado para gerar perfis populacionais imparciais ou para inferência estatística.	EDA; geração de hipóteses; aceleração do processo de descoberta de insights; cenários de classificação	Focado no perfilamento de múltiplas colunas, especificamente para a detecção de outliers, anomalias e a exploração de clusters de dados incomuns.
Bias Correction Sampling	Corrige o viés de seleção em uma amostra de big data não probabilística por meio de subamostragem usando pesos derivados de dados auxiliares externos.	Permite inferências estatísticas válidas a partir de dados com viés ; reduz o erro de estimativa.	Depende da qualidade das informações auxiliares ; pode ser complexo de implementar.	Análise de big data para inferência populacional, especialmente quando os dados são de auto-seleção (ex: redes sociais, web logs).	Uma etapa de pré-processamento antes do perfilamento, usada para corrigir amostras de fontes com viés de seleção (ex: redes sociais) e permitir um perfilamento generalizável.
Imbalanced Data Sampling	Técnicas de reamostragem (oversampling, undersampling) para balancear a distribuição de classes em um dataset.	Melhora a capacidade do modelo de aprender com a classe minoritária.	Computacionalmente complexo; oversampling pode levar a overfitting; undersampling pode descartar dados importantes.	Problemas de classificação com dados desbalanceados, como detecção de fraude, diagnóstico médico e detecção de intrusão.	Usado antes do perfilamento para balancear classes. Essencial para analisar corretamente as características da classe minoritária (ex: perfil de transações fraudulentas).
Non probability sampling	Seleção baseada em que indivíduos são selecionados da população sem o princípio da randomização.	Prático, rápido e de baixo custo, ideal para estudos exploratórios.	Alto risco de viés, pois a amostra pode não representar a população; limita a generalização dos resultados.	Pesquisas exploratórias, estudos com populações de difícil acesso ou quando a aleatoriedade é inviável por restrições de tempo/recursos.	Inadequado para o Data Profiling formal devido ao alto risco de viés. Pode ser usado para geração de hipóteses informais em fases muito iniciais da exploração.

Fonte: O autor (2025)

6. Ameaças à Validade

Em revisões sistemáticas da literatura (RSLs), as ameaças à validade representam um desafio constante, especialmente no campo da tecnologia e inovação, que é caracterizado por mudanças e evoluções contínuas. Diante disso, é fundamental analisar as principais ameaças a este estudo, levando em conta os elementos que podem impactar a abrangência, a exatidão e a pertinência dos achados.

Um desses elementos é a presença de um único revisor, logo, para mitigar vieses e assegurar um maior rigor científico, diversas precauções foram tomadas no planejamento e na execução da pesquisa, como os critérios de inclusão, exclusão e qualidade. Ademais, uma diversidade de contextos dos estudos analisados pode limitar a generalização dos resultados, assim, para mitigar essa restrição e aumentar a representatividade da revisão, foram selecionadas publicações que abrangem múltiplas perspectivas, pesquisas e análises de métodos de amostragem.

A abrangência deste estudo foi restringida ao uso exclusivo do Google Scholar e embora ela seja uma ferramenta vasta, a inclusão de bases de dados adicionais como IEEE Xplore e ACM Digital Library teria proporcionado uma perspectiva mais rica sobre o tema. A decisão de priorizar o Google Scholar justifica-se pela vasta quantidade de artigos disponíveis, pela facilidade de uso e por centralizar o conteúdo de outras bases de dados. Contudo, reconhecemos que essa escolha introduz um viés, pois apesar de suas vantagens, o Google Scholar apresenta desvantagens como a heterogeneidade, a opacidade nos critérios de indexação, e não permitir refinamentos rigorosos por metadados, área ou tipo de publicação.

Além disso, a validade da seleção dos estudos foi assegurada por meio de testes preliminares e critérios de inclusão e exclusão objetivos, enquanto a consistência dos resultados foi garantida pela padronização da extração de dados, com a exportação do Google Scholar, bem como a limitação de artigos com pontuação de qualidade superior a 80%, assegurando a robustez do conjunto amostral de acordo com os critérios de qualidade. Todo esse processo seguindo o protocolo definido por Kitchenham & Charters (2007).

Outro viés relevante é a temporalidade e a aplicação dos estudos. A maioria concentrou-se nos anos de 2024 e 2025, o que reforça a atualidade e relevância do tema. No entanto, isso também sugere que o campo ainda está em desenvolvimento

e pode passar por mudanças significativas nos próximos anos. Consequentemente, a evolução do tema pode rapidamente tornar a Revisão Sistemática da Literatura (RSL) desatualizada, exigindo revisões contínuas à medida que novas pesquisas são publicadas para mitigar esse risco.

Por fim, como o objetivo principal deste trabalho foi realizar uma RSL para consolidar o estado da arte e contextualizar as técnicas de amostragem em Big Data, a escassez de aplicações práticas, guias ou exemplos de aplicação constitui, portanto, uma limitação. Para compensar essa lacuna, buscou-se apresentar o máximo de repertório e validações dos autores de cada técnica, visando garantir a confiabilidade e guia prático refinado pelo próprio estado da arte para suas respectivas aplicações. Adicionalmente, o trabalho oferece contribuições, perspectivas e comentários próprios sobre a usabilidade, vantagens, desvantagens, recomendações de aplicação de cada técnica e um detalhamento objetivo da lógica subjacente a elas, conforme compilado na tabela 7.

7. Conclusão e Trabalhos Futuros

O presente estudo realizou uma revisão sistemática da literatura, com o propósito de investigar a aplicação da amostragem em ambientes Big Data. Os resultados indicam que a amostragem deixou de ser uma simples tática de redução de dados e se tornou um componente integral, dinâmico e estratégico de todo o ciclo de vida dos dados, e isso se deve por causa dos inúmeros tipos e aplicações existentes. É um facilitador que pode ser essencial para manter perto do Big Data e outras áreas, como Data Profiling, aprendizado de máquina e análises interativas em ambientes onde a computação exaustiva é uma impossibilidade econômica e técnica.

O futuro da análise de big data será, sem dúvida, moldado pelo desenvolvimento contínuo de técnicas de amostragem que sejam não apenas computacionalmente eficientes e estatisticamente sólidas, mas também profundamente integradas aos objetivos analíticos do profissional. Em última análise, em uma era definida pela abundância de dados, a qualidade e a pontualidade dos insights são determinadas não pelo grande volume de dados processados, mas pela inteligência e sofisticação da estratégia de amostragem empregada.

A principal contribuição deste trabalho é a consolidação do estado da arte, oferecendo um panorama organizado que serve como um guia para pesquisadores e profissionais na seleção e aplicação de métodos de amostragem. Embora a metodologia tenha sido rigorosa, o estudo possui suas limitações conforme destacado no Capítulo 6. Sendo uma revisão sistemática da literatura, o objetivo foi alcançado, portanto, não houve aplicações ou simulações práticas. No entanto, ainda existe a importância de equilibrar a teoria e a prática, e naturalmente é necessário destacá-la.

Sendo assim, para trabalhos futuros, recomenda-se a validação prática das técnicas por meio de análises comparativas de desempenho em diferentes cenários e sistemas, como um estudo de caso ou aplicações simuladas, o que acabaria por complementar este estudo teórico com uma abordagem prática.

Sugere-se também aprofundar a pesquisa sobre a integração de métodos de amostragem com IA Generativa e Large Language Models (LLMs), um campo promissor para a geração de dados de treino e teste, conforme apresentado pelo I-Sampling. Metodologicamente, futuras revisões devem ampliar o escopo para outras bases de dados além do Google Scholar, como IEEE Xplore e ACM Digital

Library, e incorporar múltiplos revisores para aumentar a robustez e a fidedignidade dos resultados, utilizando validações cruzadas e ferramentas colaborativas, como o Rayyan que otimizam o processo de triagem por pares e facilita a análise dos resultados, evitando a permanência de um único revisor.

Por fim, é fundamental que novos estudos se aprofundem no aprimoramento das técnicas de amostragem, com a necessidade de desenvolver métodos que elevem a interpretabilidade e a confiabilidade dessas abordagens, possibilitando sua aplicação mais segura e eficaz em contextos práticos, principalmente que sejam capazes de operar nas diversas camadas e vertentes dos “Vs” do Big Data, sejam eles relacionados puramente à grandes volumes de dados ou apresentando uma intersecção com perfilamento de dados aplicados em ambientes Big Data.

REFERÊNCIAS

[1] TALEB, Ikbal; SERHANI, Mohamed Adel; DSSOULI, Rachida. Big Data Quality: A Data Quality Profiling Model. In: XIA, Y.; ZHANG, L.-J. (eds.). Services 2019. Cham: Springer, 2019. p. 61-77. (Lecture Notes in Computer Science, v. 11517).

[2] SHIVAPRASAD, Nandish. Enhancing Data Quality through Automated Data Profiling. International Journal for Research Publication and Seminar, v. 15, n. 4, p. 108-117, out./dez. 2024.

[3] COUTO, Júlia Colleoni et al. New trends in big data profiling. jul. 2022.

[4] DAI, Wei et al. Data Profiling Technology of Data Governance Regarding Big Data: Review and Rethinking. In: Advances in Intelligent Systems and Computing. Cham: Springer, 2016. p. 439-450.

[5] SAMPAIO, Sandra de F. Mendes; DONG, Chao; SAMPAIO, Pedro. DQ2S – A framework for data quality-aware information management. Expert Systems with Applications, v. 42, n. 21, p. 8304-8326, 2015.

[6] LIU, Zhicheng; ZHANG, Aoqian. Sampling for Big Data Profiling: A Survey. IEEE Access, v. 8, p. 72713-72726, 2020.

[7] AZEROUAL, Otmane; SAAKE, Gunter; SCHALLEHN, Eike. Analyzing data quality issues in research information systems via data profiling. International Journal of Information Management, v. 41, p. 50-56, 2018.

[8] ELBAGHAZAOUI, Bahaa Eddine; AMNAI, Mohamed; SEMMOURI, Abdellatif. A Survey of Big Data Profiling: State-of-the-Art, Use Cases and Challenges. In: GHERABI, N.; KACPRZYK, J. (eds.). Intelligent Systems in Big Data, Semantic Web and Machine Learning. Cham: Springer Nature Switzerland AG, 2021. p. 111-122. (Advances in Intelligent Systems and Computing, v. 1344).

[9] LIM, Sungjoon et al. Data Profiling Procedure to Assess Data Quality. ICIC Express Letters Part B: Applications, v. 11, n. 2, p. 145-149, fev. 2020.

[10] JANG, Won-Jung et al. A Study on Data Profiling: Focusing on Attribute Value Quality Index. Applied Sciences, v. 9, n. 23, art. 5054, nov. 2019.

[11] RANGINENI, Sandeep et al. A Review on Enhancing Data Quality for Optimal Data Analytics Performance. International Journal of Computer Sciences and Engineering, v. 11, n. 10, p. 51-58, out. 2023.

[12] ABUDUZZAMAN, S. M.; HASAN, Piam Emrul. A Brief Review of Papers Related to Data Profiling, Measurement and Monitoring Tools. 2022.

[13] CANBEK, Gürol; SAGIROGLU, Seref; TEMIZEL, Tugba Taskaya. New Techniques in Profiling Big Datasets for Machine Learning with A Concise Review of Android Mobile Malware Datasets. In: International Congress on Big Data, Deep

Learning and Fighting Cyber Terrorism, 2018, Ankara. Anais... Ankara: IEEE, 2018. p. 117-121.

[14] JANG, Won-Jung et al. A Study on Data Profiling Based on the Statistical Analysis for Big Data Quality Diagnosis. International Journal of Advanced Science and Technology, v. 117, p. 77-88, 2018.

[15] NIKOLAKOPOULOS, Anastasios et al. Scalable Data Profiling for Quality Analytics Extraction. [s.l.]: Springer, 2023.

[16] LIU, Zhicheng; ZHANG, Aoqian. A Survey on Sampling and Profiling over Big Data. 2020. Technical Report (IEEE / arXiv), 2020.

[17] CLEMENTE, Fabiana et al. ydata-profiling: Accelerating data-centric AI with high-quality data. Neurocomputing, v. 554, art. 126585, 2023.

[18] JUDDOO, Suraj. Overview of data quality challenges in the context of Big Data. In: International Conference on Computing, Communication and Security (ICCCS), 2015, Pamplemousses. Anais... Pamplemousses: IEEE, 2015. p. 1-9.

[19] MADDALI, Raghavender. AI-Driven Data Profiling And Quality Assurance In Large-Scale Data Warehouses. International Journal of Research and Analytical Reviews, v. 10, n. 1, p. 329-340, mar. 2023.

[20] BABU, K. Makesh; KUMAR, K. Mohan. Validating A Big Data Environment Using Various Data Profiling Analysis. International Journal of Scientific & Technology Research, v. 8, n. 10, p. 2464-2467, out. 2019.

[21] NAUMANN, Felix. Big Data Profiling. Apresentação de slides. Dresden, mar. 2014.

[22] MAHMUD, Mohammad Sultan et al. A Survey of Data Partitioning and Sampling Methods to Support Big Data Analysis. Big Data Mining and Analytics, v. 3, n. 2, p. 85-101, jun. 2020.

[23] ANUPAMA, C. G.; LAKSHMI, C. A Comprehensive Review on Data Partitioning and Sampling Techniques for Processing Big Data. In: 2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), 2022. Anais... [s.l.]: IEEE, 2022. p. 1-6.

[24] CAI, Yongda et al. CDFRS: A scalable sampling approach for efficient big data analysis. Information Processing and Management, v. 61, art. 103746, 2024.

[25] DJOUZI, Kheyreddine; BEGHDAD-BEY, Kadda; AMAMRA, Abdenour. A new adaptive sampling algorithm for big data classification. Journal of Computational Science, v. 61, art. 101653, 2022.

[26] DJOUZI, Kheyreddine; BEGHDAD-BEY, Kadda; AMAMRA, Abdenour. Big Data Sampling Techniques: A State-of-the-Art Survey. In: International Conference on Artificial Intelligence and Its Applications, 2023, [s.l.]. Anais... [s.l.]: s.n., 2023.

[27] HE, Yulin et al. I-sampling: A New Block-Based Sampling Method for Large-Scale Dataset. In: 2017 IEEE 6th International Congress on Big Data (BigData Congress), 2017. Anais... [s.l.]: IEEE, 2017. p. 360-367.

[28] KHOEI, Tala Talaei; SINGH, Aditi. Data reduction in big data: a survey of methods, challenges and future directions. International Journal of Data Science and Analytics, 2024. Publicado online em 10 de julho de 2024.

[29] KIM, Jae Kwang; WANG, Zhonglei. Sampling Techniques for Big Data Analysis. International Statistical Review, v. 87, suppl. 1, p. S177-S191, 2019.

[30] LIU, Zhicheng; ZHANG, Aoqian. A Survey on Sampling and Profiling over Big Data. arXiv Technical Report, 2020.

[31] MADDALI, Mohammad Sultan et al. A Survey of Data Partitioning and Sampling Methods to Support Big Data Analysis. Big Data Mining and Analytics, v. 3, n. 2, p. 85-101, jun. 2020.

[32] ROJAS, Julian A. Ramos et al. Sampling Techniques to Improve Big Data Exploration. In: 2017 IEEE Symposium on Large Data Analysis and Visualization (LDAV), 2017, Phoenix. Anais... Phoenix: IEEE, 2017. p. 26-35.

[33] SALLOUM, Salman; HUANG, Joshua Zhexue; HE, Yulin. Exploring and cleaning big data with random sample data blocks. Journal of Big Data, v. 6, n. 45, 2019.

[34] TALEB, Ikbal et al. Big Data Quality: A Quality Dimensions Evaluation. In: IEEE International Conference on Big Data, 2016. Anais... [s.l.]: IEEE, 2016.

[35] ZAYED, Mohmmed Mohammed; BA-ALWI, Fadl Mutaher. Big Data Formation, Reduction, and Its Impact on Sampling: A survey. Sana'a University Journal of Applied Sciences and Technology, v. 3, n. 1, p. 597-603, 2025.

[36] ZHU, Bailin; WANG, Hongliang; FAN, Mi. Constructing small sample datasets with game mixed sampling and improved genetic algorithm. The Journal of Supercomputing, v. 80, p. 20891-20922, 2024.

[37] MOHER, David et al. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Medicine, v. 6, n. 7, p. e1000097, 2009.

[38] KITCHENHAM, Barbara; CHARTERS, Stuart. Guidelines for performing Systematic Literature Reviews in Software Engineering. EBSE Technical Report EBSE-2007-01. Keele University; University of Durham, 2007.