DOI: 10.5748/19CONTECSI/PSE/DSC/7016

TÉCNICAS DE ANONIMIZAÇÃO DE GRANDES VOLUMES DE DADOS (BIG DATA) EM BLOCKCHAIN

Robson Silva ; https://orcid.org/0000-0002-7168-1697
Universidade Salvador (UNIFACS) - Programa de Pós-graduação

Técnicas de Anonimização de Grandes Volumes de dados (Big Data) em *Blockchain*

XXXXX[1], YYYYY[1]
[1] rrrrrrr@aaaa.cccc; ppppp@uuuuu.xx
UUUUUUUU

Abstract

Technological advancement has made relationships between companies and people increasingly data-oriented; the processing, analysis and storage of data require technologies that enable short-time responses for decision making. With the exponential growth of data generated by devices, social networks and technological integration, the new era called Big Data emerges, which highlights the volume, variety and speed of the extensive mass of data. However, new challenges also arise amid the technical limitations applied to ensure safety in the use of data without compromising exposing sensitive information, some solutions have been proposed in order to mitigate this gap in government, bank and political projects. Cloud platforms where most big data sets are stored and scalable data processing tools using cloud infrastructure have become increasingly attractive to support big data mining and analytics applications. Despite clouds having many salient features such as cost effectiveness and high scalability, privacy concerns are one of the main obstacles to the adoption of public cloud capabilities in many privacy sensitive applications in industries such as healthcare, finance and defense. The barriers faced in analyzing, integrating and sustaining data anonymization have motivated the development of strategies to ensure people's success and safety. Anonymous data security is crucial so that decisions based on the results achieved are relevant. In this one, comprehensive research on data anonymization techniques will be presented, for this, it was necessary to understand the state of the art and identify the anonymization techniques of large volumes of data in blockchain. To achieve the objective was to understand the application of data anonymization techniques through an approach composed of the literature review of the main techniques of anonymization of large volumes of data in Public Blockchain. Subsequently, these techniques are accounted for according to the literature. The result of this work reveals that the two main techniques used when compared in public blockchains are widely accepted in the literature, and are even used in works as proposals for new solutions.

Keywords*: smart contract, big data, blockchain, data anonymization, data security.*

Resumo

O avanço tecnológico tornou as relações entre empresas e pessoas cada vez mais orientada a dados, o processamento, análise e armazenados dos dados exigem tecnologias que possibilitem respostas em curto período para tomada de decisão. Com o crescimento exponencial dos dados gerados pelos dispositivos, redes sociais e integração tecnológica surge a nova era denominada *Big Data* no qual se destaca o volume, variedade e velocidade da extensa massa dados. No entanto, novos desafios também surgem em meio as limitações técnicas aplicadas para garantir a segurança na utilização de dados sem expor as informações de caráter sensível, algumas soluções foram propostas a fim de mitigar essa lacuna em projetos governamentais, bancário e político. As plataformas de nuvem onde a maioria dos conjuntos de *big data* são armazenados e as ferramentas de processamento de dados escaláveis usando infraestrutura de nuvem tornaram-se cada vez mais atrativas para dar suporte a aplicativos de mineração e análise de big data. Apesar das nuvens terem muitos recursos, como custo-benefício e alta escalabilidade, as preocupações com a privacidade são um dos principais obstáculos para a adoção dos recursos da nuvem pública em muitas aplicações sensíveis à privacidade em setores como saúde, finanças e defesa. As barreiras enfrentadas em analisar, integrar e sustentar a anonimização de dados tem motivado o desenvolvimento de estratégias para garantir o sucesso e a segurança das pessoas. A segurança de dados de forma anônima é crucial para que as decisões com base nos resultados alcançados sejam relevantes. Neste artigo será apresentado uma pesquisa sobre técnicas de anonimização de dados, para isso, é necessário compreender o estado da arte e identificar as técnicas de anonimização de grandes volumes de dados em *blockchain*, uma vez que esta tecnologia permite registro e transações com dados com segurança, tornando-os imutáveis. Para atingir o objetivo deste trabalho foi necessário entender a aplicação de técnicas de anonimização de grandes volumes de dados em *blokchain* pública, por meio de uma revisão de literatura. O resultado deste trabalho revela que as duas principais técnicas utilizadas quando comparadas na *blokchain* pública tem ampla aceitação na literatura, inclusive são utilizadas em trabalhos como propostas de novas soluções.

Palavras Chaves: *big data, blockchain, anonimização de dados, segurança de dados*

1 Introdução

Com os progressos tecnológicos grandes quantidades de dados são coletados diariamente sobre as pessoas, esses dados vêm aumentando em termos de complexidade, diversidade e atualidade [1], as organizações governamentais e privadas são alguns exemplos das que utilizam mecanismos e ferramentas como meio de capitação de dados. As atividades diárias de um indivíduo proporcionam subsídios para os mecanismos se apropriarem de dados, algumas das fontes de capitação são sistemas de compras online, sensores, transações bancárias, prontuários médicos. De acordo com [1], aponta que a quantidade de dados produzidas entre 2006 e 2010 cresceu de 166 Exabytes para 988 Exabytes, com estimativa que até o ano de 2020 esse volume aumente em torno de 40.000 Exabytes ou 40 Zettabytes [2]. Essa tendência marcou o início de um novo paradigma denominado *Big Data* que gerou novos desafios em termos de velocidade, variedade, variabilidade e complexidade [2] [4], em contrapartida [3], afirma ser um conceito abstrato, porém, caracteriza como os volumes de dados que não é possível ser gerenciados e processados por ferramentas tradicionais de software ou hardware dentro de um tempo tolerável.

Os serviços médicos concentram uma enorme quantidade de dados da saúde dos usuários, afirma [8]. Anualmente a Organização Mundial de Saúde com base nos dados coletados publica relatórios resultados estatísticos sobre a saúde da população mundial. A busca pela posse de dados se tornou uma moeda com grande valor de mercado, tendo em vista que a análise desses dados pode favorecer a competitividade, bem como a comercialização de medicamentos e/ou fornecimento de serviços, o anuário estatístico do mercado farmacêutico detalha o comportamento do setor que em 2017 movimentou mais de 69,5 bilhões com vendas de medicamentos [6].

No âmbito da pesquisa cientifica houve muitos avanços no campo da saúde, sendo capaz de empregar os resultados para mapear regiões, classificar grupos de riscos, previsões e inferências de doenças [6]. Os dados de saúde gerados pelos usuários são cada vez mais compartilhados, divulgados e publicados para pesquisa [6,7], negócios, aplicações governamentais [9, 10] e outros cenários [8, 6]. Além do valor na utilização, os dados trazem detalhes sensíveis que podem gerar significativos riscos à privacidade dos indivíduos. O termo privacidade pode ter diferentes abordagens e definições, uma das mais conservadora é descrito por [9]. A anonimização do ponto de vista de [14], é uma solução prática para preservar a privacidade do usuário na publicação de dados. Os proprietários de dados, como hospitais, bancos, provedores de serviços de redes sociais e grande instituições, anonimizam os dados de seus usuários antes de publicá-los para proteger a privacidade, enquanto os dados anônimos permanecem úteis para consumidores legítimos de informações. Por outro lado, [12] destaca que a anonimização de dados é a técnica em que as informações que divulgam a identidade são removidas dos conjuntos de dados, ou seja, os dados sensíveis não são identificáveis, embora seu formato e tipo de dados sejam preservados.

Inúmeros modelos, algoritmos, estruturas e protótipos de anonimização foram propostos/desenvolvidos para publicação de dados com preservação de privacidade. Esses modelos/algoritmos anonimizam os dados dos usuários, principalmente na forma de tabelas ou gráficos, dependendo dos proprietários dos dados [10]. A maioria das organizações, como hospitais, bancos, seguradoras e supermercados, coleta dados relevantes de clientes/assinantes para melhorar a qualidade do serviço. As redes sociais permitem que usuários possam interagir de maneira que dados sejam compartilhados e acessados, esses dados são utilizados por empresas para campanhas de publicidade, recomendação de produtos, serviços e classificação do perfil de consumo. Entretanto, alguns dados de caráter sensível, em que segundo a legislação europeia, conhecida como Regulamento Geral de Proteção de Dados [11] e a Lei Geral de Proteção de Dados [12], caracteriza os dados sensíveis sendo aqueles que podem identificar a pessoa, tais como: origem racial ou étnica, genética, crenças religiosas ou filosóficas, dados

relativos ao estado de saúde ou à vida sexual e/ou orientação sexual. Esses dados coletados geralmente contêm informações sobre as atividades do usuário, dados demográficos, finanças, hobbies, localização, percepções, interesses, preferências, visões políticas e religiosas, comunidades online e opiniões [10].

Destarte, será apresentado neste trabalho uma revisão da literatura com o objetivo de identificar pesquisas que lidam com o tema de anonimização de grandes volumes de dados na *blockchain*.

Este trabalho está estruturado da seguinte forma: A Seção 2 apresenta a metodologia utilizada para a realização da revisão da literatura; A Seção 3 discute os principais conceitos relacionados à anonimização de dados e blockchain; Na Seção 4 se discute as técnicas de anonimização de dados em Big Data; A Seção 5 apresenta os resultados desta revisão da literatura; Na Seção 6 são apresentadas as considerações finais deste trabalho e possíveis trabalhos futuros e a Seção 7 Lista os trabalhos que fazem parte deste estudo.

2 Metodologia

A revisão de literatura foi realizada seguindo os critérios descritos em [24] e aplicado no contexto da Anonimização de Dados, conforme ilustrado na Figura 1,. O escopo foi definido com base na seguinte questão de pesquisa: "Quais são as técnicas de anonimização de grandes volumes de dados (*big data*) em *blockchain*?". Esta questão de pesquisa foi organizada em dois critérios: (1) Artigos publicados sobre anonimização de grandes volumes de dados; (2) Técnicas de anonimização aplicadas em *blockchain*.

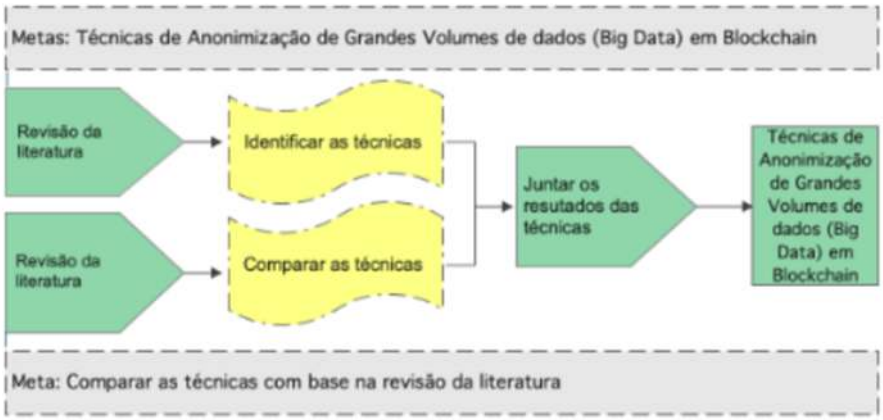

Figura 1 – Estratégia de pesquisa

Antes de iniciar a execução do protocolo, foi definida uma *string* de busca inicial, que a princípio considerou os critérios citados previamente: (*big data* ou anonimização de dados ou *blockchain* ou *smart contract*). A aplicação da *string* em algumas bibliotecas digitais resultou em muitos estudos encontrados. Assim, foi necessário ajustar a *string* de busca e, consequentemente, minimizar o risco de algum trabalho relevante para esta busca não ser encontrado. A nova *string* de pesquisa estava relacionada apenas a artigos que relatavam algo sobre anonimização de dados e *blockchain*. Com isso, a *string* de pesquisa considerada foi: *((big data and data anonymization and blockchian) and (smart contract and data security)).* A formatação das *strings* de pesquisa nos repositórios foi estruturada de maneira que fossem realizadas nos termos inseridos.

Critérios de inclusão e exclusão também foram definidos. Os critérios de inclusão foram: (i) artigos publicados descrevendo técnica de anonimização de dados em *blockchain*; (ii) artigos que abordam o mesmo estudo, apenas o mais recente será incluído; (iii) o ano de publicação entre 2011 e 2021, conforme a Figura 2. Em contrapartida, os critérios de exclusão foram: (i) estudos com informações insuficientes sobre anonimização de dados em *blockchain*;

(ii) artigos disponíveis como resumos ou apresentações em PowerPoint. As bibliotecas digitais consideradas neste trabalho foram: Google Schoolar, Scopus, IEEE Xplorer.

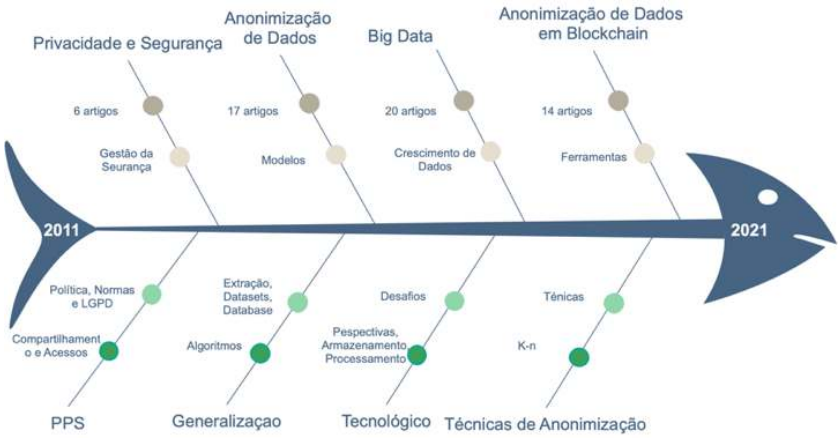

Figura 2 – Número de artigo com base *string* de busca

O processo de seleção dos estudos foi dividido em três etapas. A primeira etapa representa o número total de estudos retornados pela execução da *string*. Foi realizado entre os anos de 2011 e 2021 e resultou em 348 estudos para avaliação. Durante esse processo, foram aplicados os critérios de inclusão e exclusão que resultou em 24 artigos selecionados para leitura e análise. A partir da análise de cada estudo, foi possível responder à questão de pesquisa definida no escopo.

3 Anonimização de dados e *Big data*

A anonimização de dados permite a divulgação de informações inteiras que são úteis para consultas e análises, mantendo a privacidade de dados confidenciais contra diversos tipos de ataques. O trabalho de [16], apresentou uma revisão das técnicas de anonimização empregadas em dados, ataques e riscos de privacidade. O Privacy Policy Manager (PPM) [8] gerencia as configurações de privacidade de um indivíduo e fornece informações para controlar os fluxos de dados privados de acordo com essas configurações de privacidade. No processo de anonimização um data broker gera um conjunto de dados anônimos na primeira etapa do protocolo de transferência de dados.

Antes de enviá-lo, o corretor de dados deve confirmar o risco de privacidade do conjunto de dados anônimos. Na avaliação de riscos de privacidade, a diretriz básica é o k-anonimato; o mínimo k deveria ser definido como um regulamento de distribuição de dados e o data broker geraria um conjunto de dados anôncimos para atender ao regulamento. Segundo [13] as técnicas predominantes para anonimização de dados podem ser categorizadas da seguinte forma:

Natureza dos dados: Várias técnicas são sugeridas para:

i. Natureza dos dados: Várias técnicas são sugeridas para:

- *Dados tabulares*: Inclui informações sobre indivíduos, seus quase-identificadores associados, como idade, sexo, código PIN etc.; e informações confidenciais associadas, como salário ou doenças.

- *Conjuntos de dados de itens*: denota dados transacionais que associam os indivíduos com seus conjuntos de itens comprados em transações.
- *Dados gráficos*: denotam associações sensíveis do indivíduo, como conectividade de redes sociais de indivíduos.

ii. Abordagens de anonimização: A seguinte variedade de abordagens já é proposta:

- *Generalização*: O atributo como idade é reforçado em conjuntos de dados. Por exemplo, a idade é generalizada em faixas etárias.
- *Supressão*: O atributo como gênero é destacado do conjunto de dados completo.
- *Perturbação*: Adição de ruído aos atributos, como salário no conjunto de dados.
- *Permutação*: As ligações sensíveis entre instâncias são trocadas como a compra de medicamentos por um indivíduo.

iii. Modelos: Associa o conjunto de dados.

- *k-anonimato*: indica que cada indivíduo no conjunto de dados deve ser idêntico a k−1 outros indivíduos.
- *L-diversidade*: Busca a diversidade adequada para os atributos sensíveis relacionados aos indivíduos.

3.1 Ecossistema *Big Data*

A estrutura do ecossistema de *big data* deve ser escalável, adaptativa e possuir recursos inteligentes de processamento de dados. O *Hadoop* é uma estrutura para armazenamento e processamento em larga escala de conjuntos de dados [16]. Um ecossistema de redes inteligentes é descrito por [18] e propõem a utilização do Haddop *Big data* Lake para servir de armazenamento de redes inteligentes de dados incorporando medidor inteligente, imagem e dados de vídeo. A Figura 3 de maneira resumida descreve a composição do sistema Hadoop.

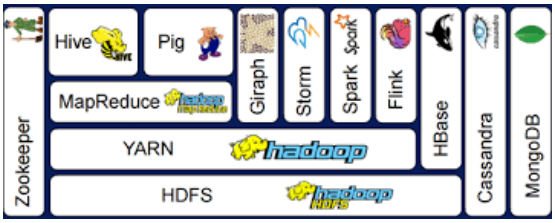

Figura 3 – Ecossistema Hadoop

3.2 Análise de *Big Data*

O processo de examinar grandes volumes de dados para identificar tendências e relacionamentos usando tecnologias avançadas para desenvolver modelos de predição e algoritmos estatísticos complexos, e que auxiliem nas decisões de negócios, é conhecido como análise de *big data*. O processamento e análise de *big data* pode ser atingindo se o ambiente possuir dispositivos e recursos de software inteligentes para coletar dados de várias fontes, processá-los automaticamente em informações significativas para fornecer novos *insights* na tomada de decisões. As análises aplicadas à *big data* podem incluir mineração de dados, aprendizado de máquina, análise de redes sociais e visualizações de dados.

O resultado das análises ajuda a identificar oportunidades de redução de custos, criação de oportunidades, integração de dados e possíveis vantagens competitivas. As análises utilizam níveis descritivos, prescritivos ou preditivos. A prescritiva pode recomendar algo (e.g. otimização de algum recurso), a predição é capaz de aplicar algoritmos de classificação, e a análise descritiva utiliza-se de técnicas de relação [17]. Portanto, a análise de *big data* pode ser aplicada em vários domínios de conhecimento para anonimização de dados. Os possíveis tipos de dados encontrados em um ambiente de *big data* podem ser classificados como:

- *Dados estruturados*: A origem das informações é ordenada sobre uma estrutura de dados organizacional, a exemplo, dados oriundos de vendas, formulários em lote, estruturas de sistema de gestão empresarial, estruturas de empreendimentos ampliadas;
- *Dados não estruturados*: São dados que não possuem uma estrutura lógica organizacional e podem variar de acordo as fontes, destaca-se alguns exemplos que são as redes online, mensagens de texto, mensagens, conexões, gravações, imagens e documentos sonoros;
- *Dados semiestruturados*: São dados em que a organização estrutural contém elementos que podem dividir os dados em hierarquias separadas [18].

A Figura 4 ilustra os tipos de organização de dados.

Figura 4 – Tipo de dados

3.3 Tecnologia *Blockchain*

A tecnologia *blockchain* atraiu significativa atenção de pesquisadores, engenheiros, economistas e empreendedores nos últimos anos. É uma tecnologia baseada em transações descentralizadas, imutáveis e tolerantes a falhas. Atualmente a *blockchain* é considerada a espinha dorsal da segurança na internet. Na origem da *blockchain*, de acordo com [5], está o protocolo do Bitcoin, proposto por Satoshi Nakamoto Nakamoto 2008, que entrou em operação em 2009. O artigo propõe uma rede P2P onde transações com a criptomoeda bitcoin, propostas por clientes, são recebidas por servidores que irão decidir, através de um protocolo de consenso bizantino a base de desafios criptográficos, sobre a ordem em que serão realizadas e armazenadas permanentemente numa cadeia de blocos (blockchain), replicada em cada servidor [21].

Neste contexto, a *blockchain* é um livro-razão livremente replicado e armazenado em diferentes nós da rede Bitcoin, tornando o Bitcoin um sistema completamente distribuído [20]. Segundo a definição de [21], a *blockchain* implementa uma máquina de estados replicada para a manutenção consistente de um estado global compartilhado por um conjunto de pares distribuídos numa rede P2P. As transações são inseridas em determinado período e uma vez que todas as transações recém criadas do sistema são coletadas, elas são compiladas juntas em uma estrutura de dados, chamada de blocos. Segundo [20], esse bloco é então inserido no topo

da *blockchain*. Além disso, cada bloco é identificado por seu hash criptográfico e conectado com outro bloco para fazer cadeia [7].

Uma transação é chamada de "transação confirmada" quando o bloco que contém essa transação é inserido na *blockchain*. É possível verificar uma transação confirmada para evitar ataques. Alguns problemas que persiste na *blockchain* é o consumo energético usado para processar as transações e o tempo. Um mecanismo geral de operação de *blockchain* é explicado na Figura 5, de maneira que a cadeia de blocos é composta pelo *header*, *hash*[1] do bloco anterior e transações.

Figura 5 – Blockchain e o hash do bloco

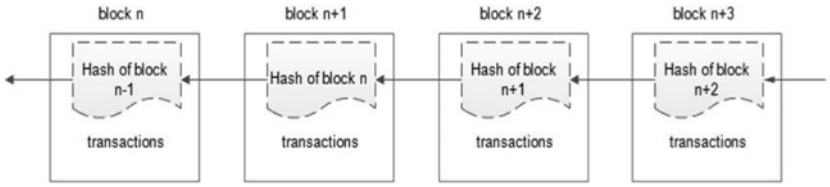

Figura 5 – Blockchain e o hash do bloco [7]

4. Técnicas de Anonimização de Dados

As seções seguintes discutem as técnicas de anonimização de dados em *Big Data* encontradas com maior frequência na revisão da literatura, são elas *Map* e *Reduce* e RDD.

4.1 Técnicas de Análise de Dados *Map* e *Reduce*

Map e *Reduce* (MPR) é um modelo de programação para processamento em paralelo de grandes volumes de dados distribuídos em clusters. Também conhecido como paradigma de programação, em que uma aplicação MPR consiste em dividir e processar conjuntos de dados com o uso das duas principais funções *Map* e *Reduce*, não sendo necessário realizar nenhum tipo de programação extra para garantir que os processos serão executados de forma paralela.

Considerado uma abordagem para resolver problemas envolvendo análise de dados em larga escala. A Figura 6 ilustra as etapas que envolve as funções *Map* e *Reduce* durante o processo de análise de grandes volumes de dados.

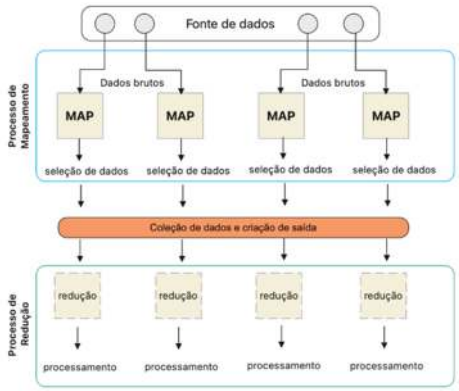

Figura 6 – Estrutura *Map-Reduce*

[1] As funções hash criptográficas são algoritmos matemáticos usados para mapear dados de qualquer tamanho para uma sequência de bits de tamanho fixo, e.g. o SHA-256 é um algoritmo de hash seguro de 256 bits usado para proteção criptográfica.

4.2 Técnicas de Análise de Dados RDD

Durante a revisão da literatura foram identificados vários trabalhos que fortalecem o uso da *blockchain* para anonimização de grandes volumes de dados, porém, uma grande parte tem como *insight* a utilização do ecossistema *Hadoop* como principal técnica de processamento de grandes volumes de dados, aliado a técnica de *Map-Reduce* para construção de *framework*. Entretanto, o ecossistema *Apache Spark* foi citado em diversos estudos comparativos, nos quais os resultados foram satisfatórios no contexto da *blockchain*. Podemos verificar que os estudos analíticos realizado por [15], [19], [21] e [22], trazem um comparativo de ferramentas para *Big Data Analytics* para anonimização.

Um outro trabalho relevante realizado por [23], afirma que *Hadoop* é a principal ferramenta para armazenar e analisar dados com computação de alto desempenho, estável e confiável para os diferentes tipos de dados, o autor [23] ressalta que a união com o *Spark* possui um modelo de programação conhecido como *Resilient Distributed Datasets* (RDD) possibilitando menos sobrecarga para lidar com a alta velocidade e o volume de dados. Segundo [22], o ecossistema *Apache Spark* é uma ferramenta capaz de processar e analisar dados de forma paralela e distribuída, combinados a SQL, *streaming* e análises complexas. A Figura 7 ilustra o resumo do processo de RDD.

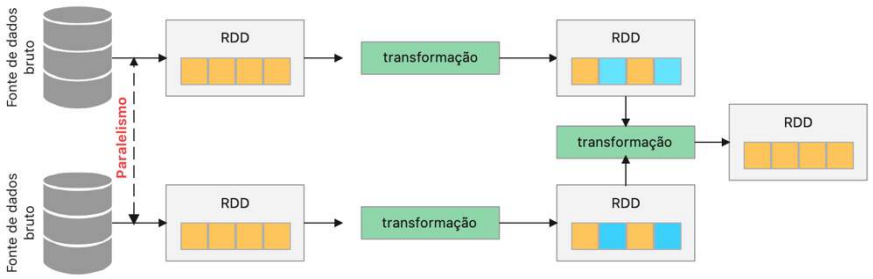

Figura 7 – Processo de RDD

As duas principais técnicas durante o processo de análise aplicam abordagem diferentes nas quais à medida que os resultados de cada etapa são processados novas etapas subjacentes são aplicadas. O trabalho [20] faz um comparativo das principais técnicas de anonimização de dados e propõem uma plataforma distribuída para anonimização de dados. Já o trabalho de [10] também corrobora em fazer uma comparação detalhada das diferentes soluções técnicas para anonimização de dados relacionais.

Alguns trabalhos trazem propostas em que a aplicação das técnicas utilize a *blockchain* com meio de armazenar as características dos tipos de dados, outros autores propõem usar amostras dos dados e convertê-los em não estruturados como forma adicional na criação do *hash*. O algoritmo hash é conhecido como uma função matemática que é capaz de converte uma entrada de comprimento arbitrário em uma saída criptografada de comprimento fixo e único.

5 Resultados

Ao estudar as priniciapais técnicas, os trabalhos selecionados mostram que em sua grande parte 18 (dezoito) artigos aplicam o RDD para análises e anonimização de grandes volumes de dados com *blockchain*. A Figura 8 mostra os resultados do número de artigos que

aplicam as técnicas RDD, MPR e outras, bem como, o número de citações com base na técnica. O baixo poder computacional exigido em paralelo ao alto poder de processamento e escalabilidade flexível, justifica a ampla adoção do Spark, i.e. RDD. Em seguida vem MPR com 5 (cinco) artigos e, por fim, outras técnicas pouco conhecidas. No quisito número de citações o RDD também supera, com 17 (dezessete), em seguinda 7 (sete) do MPR e uma para outras. Apesar deste trabalho de investigação mostrar que o Spark é amplamente aplicado, o seu concorrente direto, o MPR, tem se mostrado promissor ao longo do tempo.

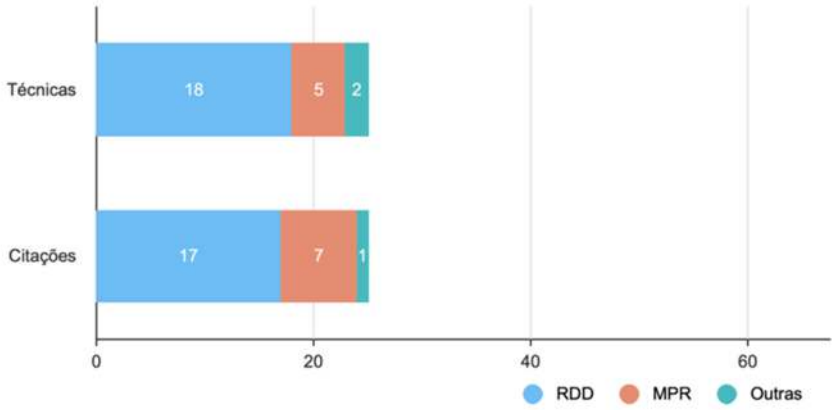

Figura 8 – Técnicas Anonimização de dados em *Blockchain*

6 Conclusão

Este artigo discutiu as principais técnicas de anonimização de conjuntos de dados em *blockchain* encontradas na literatura. Os artigos encontrados nesta pesquisa bibliográfica indicam que RDD e *Map-Reduce* com o algoritmo k-anônimo mostram-se promissor na garantia da segurança de dados. É possível inferir que estudos sobre anonimização de dados em *blockchain* ainda se encontram em estágio embrionário e que muitos estudos ainda necessitam serem explorados à medida que novos desenvolvimentos continuam para o aperfeiçoamento da segurança de grandes volumes dados e oportunidades da aplicação da *blockchain*.

Ficou evidenciado que as técnicas de anonimização de grandes volumes de dados no contexto de *blockchain* é um campo ainda pouco explorado e as ferramentas existentes são limitadas. No entanto, a quantidade de dados vem aumentando em diversos domínios e com velocidade, o que poderá tornar as principais técnicas de anonimização atualmente usadas, *Map Reduce* e RDD, inúteis ao longo tempo. Portanto, há necessidade de ampliar os estudos de maneira que seja capaz de visualizar e explorar a integração de técnicas de anonimização para *big data* no contexto da *blockchain*.

7 Referências Bibliográficas

[1] Patil, H. K., & Seshadri, R. (2014, June). Big data security and privacy issues in healthcare. In *2014 IEEE international congress on big data* (pp. 762-765). IEEE.

[2] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

[3] Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications*, *19*(2), 171-209.

[4] Grossman, R., & Siegel, K. (2014). Organizational models for big data and analytics. *Journal of Organization Design*, *3*(1), 20-25.

[5] Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review*, 21260.

[6] Nie, L., Wang, M., Zhang, L., Yan, S., Zhang, B., & Chua, T. S. (2015). Disease inference from health-related questions via sparse deep learning. *IEEE Transactions on knowledge and Data Engineering*, *27*(8), 2107-2119.

[7] Li, Z., Barenji, A. V., & Huang, G. Q. (2018). Toward a blockchain cloud manufacturing system as a peer to peer distributed network platform. *Robotics and computer-integrated manufacturing*, *54*, 133-144.

[8] Gkoulalas-Divanis, A., Loukides, G., & Sun, J. (2014). Publishing data from electronic health records while preserving privacy: A survey of algorithms. *Journal of biomedical informatics*, *50*, 4-19.

[9] Hripcsak, G., Bloomrosen, M., FlatelyBrennan, P., Chute, C. G., Cimino, J., Detmer, D. E., ... & Wilcox, A. B. (2014). Health data use, stewardship, and governance: ongoing gaps and challenges: a report from AMIA's 2012 Health Policy Meeting. *Journal of the American Medical Informatics Association*, *21*(2), 204-211.

[10] Gavison, R. (1980). Privacy and the Limits of Law. *The Yale law journal*, *89*(3), 421-471.

[11] Majeed, A., & Lee, S. (2020). Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE access*, *9*, 8512-8545.

[12] Regulation, G. D. P. (2018). General data protection regulation (GDPR). *Intersoft Consulting, Accessed in October*, *24*(1).

[13] Schwaitzer, L., Nascimento, N., & de Souza Costa, A. (2021). Reflexões sobre a contribuição da gestão de documentos para programas de adequação à Lei Geral de Proteção de Dados Pessoais (LGPD). *Acervo*, *34*(3), 1-17.

[14] Goswami, P., & Madan, S. (2017, May). Privacy preserving data publishing and data anonymization approaches: A review. In *2017 International Conference on Computing, Communication and Automation (ICCCA)* (pp. 139-142). IEEE.

[15] Kiyomoto, S., Rahman, M. S., & Basu, A. (2017, June). On blockchain-based anonymized dataset distribution platform. In *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*(pp. 85-92). IEEE.

[16] Abawajy, J. H., Ninggal, M. I. H., & Herawan, T. (2016). Privacy preserving social network data publication. *IEEE communications surveys & tutorials*, *18*(3), 1974-1997.

[17] Verma, C., & Pandey, R. (2016, January). Big Data representation for grade analysis through Hadoop framework. In *2016 6th international conference-cloud system and big data engineering (Confluence)* (pp. 312-315). IEEE.

[18] Munshi, A. A., & Mohamed, Y. A. R. I. (2018). Data lake lambda architecture for smart grids big data analytics. *IEEE Access*, *6*, 40463-40471.

[19] Nguyen, T., Li, Z. H. O. U., Spiegler, V., Ieromonachou, P., & Lin, Y. (2018). Big data analytics in supply chain management: A state-of-the-art literature review. *Computers & Operations Research*, *98*, 254-264.

[20] Hu, E. W., Su, B., & Wang, J. (2017, June). Software performance prediction at source level. In *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)* (pp. 263-270). IEEE.

[21] Greve, F. G., Sampaio, L. S., Abijaude, J. A., Coutinho, A. C., Valcy, Í. V., & Queiroz, S. Q. (2018). Blockchain e a Revolução do Consenso sob Demanda. *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)-Minicursos*.
Greve, F. G., Sampaio, L. S., Abijaude, J. A., Coutinho, A. C., Valcy, Í. V., & Queiroz, S. Q. (2018). Blockchain e a Revolução do Consenso sob Demanda. *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)-Minicursos*.

[22] Putra, H. Y., Putra, H., & Kurniawan, N. B. (2018, October). Big data analytics algorithm, data type and tools in smart city: A systematic literature review. In *2018 International Conference on Information Technology Systems and Innovation (ICITSI)*(pp. 474-478). IEEE.

[23] Benlachmi, Y., & Hasnaoui, M. L. (2020, July). Big data and spark: Comparison with hadoop. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)* (pp. 811-817). IEEE.

[24] Marconi, M. D. A., & Lakatos, E. M. (1990). Técnicas de pesquisa. *São Paulo: Atlas*. *ISBN 85-224-3263-5*