DOI: 10.5748/20CONTECSI/PSE/DSC/7228

eLocator: e207228

PROCESSAMENTO E ANÁLISE DE DADOS DO BLOCKCHAIN PARA IDENTIFICAÇAO DE ENDEREÇOS SUSPEITOS DE TRANSAÇÕES ILÍCITAS: UMA ABORDAGEM DE BIG DATA

Hugo Martinelli Watanuki – https://orcid.org/0000-0002-1530-8054
Instituto Federal De Educação, Ciência E Tecnologia De São Paulo Campus Campinas

Bianca Maria Pedrosa
Instituto Federal De Educação, Ciência E Tecnologia De São Paulo Campus Campinas

BLOCKCHAIN DATA PROCESSING AND ANALYSIS TO IDENTIFY ADDRESSES SUSPECTED OF ILLICIT TRANSACTIONS: A BIG DATA APPROACH

ABSTRACT
The objective of this work is to propose a workflow of processing and analysis of massive volumes of data from the bitcoinblockchain in order to evaluate the similarity of transaction patterns between two addresses suspected of illicit transactions. For this purpose, a computer simulation approach was adopted. As a starting point, the state of the art of specialized literature was reviewed to identify the main big data technologies capable of contributing to the processing and analysis of blockchain data, with a focus on parallel programming and distributed file systems. Next, a computer simulation environment was developed to extract and transform binary data from the bitcoinblockchain and structure it so that the RTI (Random Time Interval) of the transaction origin addresses is obtained. The RTI of such addresses were then compared with the RTI of addresses previously associated with illicit activities in order to identify the portfolio of addresses related to individuals carrying out illicit activities using cryptocurrencies. This work is expected to contribute to greater understanding regarding the use of big data technologies for the optimized processing and analysis of blockchain data.
Keywords*: Blockchain; Data processing and analysis;Big data architecture; Illicit transactions.*

PROCESSAMENTO E ANÁLISE DE DADOS DO *BLOCKCHAIN*PARA IDENTIFICAÇAO DE ENDEREÇOS SUSPEITOS DE TRANSAÇÕES ILÍCITAS: UMA ABORDAGEM DE *BIG DATA*

RESUMO
O objetivo deste trabalho é propor um fluxo de processamento e análise de volumes massivos de dados do *blockchain* do *bitcoin* com o intuito de avaliar a similaridade de padrões de transações entre dois endereços suspeitos de transações ilícitas. Para essa finalidade, optou-se por uma abordagem de simulação computacional. Como ponto de partida, o estado da arte da literatura especializada foi revisado para identificar as principais tecnologias de *big data* capazes de contribuir com o processamento e análise de dados de *blockchain*, com enfoque em programação paralela e sistemas distribuídos de arquivos. Em seguida, foi desenvolvido um ambiente de simulação computacional para extrair e transformar os dados binários do *blockchain* do *bitcoin* e estruturá-los para que o RTI (*Random Time Interval*) dos endereços de origem das transações fossem obtidos. O RTI de tais endereços foram então comparados com o RTI de endereços previamente associados a atividades ilícitas com o intuito de identificar a carteira de endereços relacionadas com indivíduos que praticam atividades ilícitas usando criptomoedas. Espera-se que este trabalho contribua para um maior entendimento em relação ao uso de tecnologias de *big data* para o processamento e análise otimizados de dados de *blockchain*.
Palavras-chave:*Blockchain*; Processamento e análise de dados; Arquiteturas de *big data*; Transações ilícitas.

1. Introdução

Tecnologias *blockchain* podem ser definidas como redes descentralizadas nas quais seus membros têm completo acesso para monitorar todas as transações na rede usando uma abordagem *peer-to-peer*, sendo que tais redes possuem como fundamentos os princípios da imutabilidade, descentralização, segurança e consenso (Deepa et al., 2022). Embora originalmente concebidas para suportar criptomoedas, os princípios fundamentais de tais redes também as tornam atrativas para aplicações inovadoras que recentemente têm sido vistas em áreas como combate a fraudes financeiras, gestão de cadeias de abastecimento e adoção em ambiente de chão de fábrica industrial (Akcora, Dixon, Gel &Kantarcioglu, 2018; Deepa et al., 2022).

Especificamente no caso de aplicações de combate a atividades ilícitas ou fraude financeira envolvendo criptomoedas, o processamento e análise de dados de *blockchain* já permite a identificação de endereços suspeitos com base em padrões de transação comuns ao longo do tempo (Wu et al., 2020; Maheshwari et al., 2023). Contudo, se por um lado processar e extrair dados do *blockchain* tem se tornado uma atividade cada vez mais importante, por outro, coletar e processar os dados completos de um *blockchain* não é uma tarefa trivial. Dentre os principais desafios reportados até o momento pode-se citar, além do tamanho considerável dos dados, a evolução constante dos protocolos que governam o *blockchain* e o fato de seus usuários usarem técnicas de ofuscação para preservar suas privacidades. Tais elementos fazem com que tarefas aparentemente triviais, tais como contar o número de transações envolvendo um determinado usuário sejam difíceis de serem realizadas em escala comercial (Emery &Latapy, 2021). Nesse contexto, analisar os dados de *blockchain* de maneira eficiente tem representado um importante tema de pesquisa nas áreas de engenharia e ciência de dados (Akcora, Dixon, Gel &Kantarcioglu, 2018; Deepa et al., 2022; Emery &Latapy, 2021).

Dado esse contexto, atualmente, já se considera que o problema de detecção de atividades ilícitas ou fraudulentas envolvendo criptomoedas representa um problema de processamento e análise de volumes massivos de dados, ou *big data*, e recursos computacionais usuais tem-se mostrado incapazes de acompanhar as demandas de processamento de dados de *blockchain* (Zhou et al., 2020; Maheshwari et al., 2023). Por conta desse racional, uma possível abordagem para o problema consiste no uso de supercomputação ou computação paralela e distribuída, recurso que historicamente é utilizado para tratar problemas de processamento e análise de *big data* (Emery &Latapy, 2021). Alguns estudos recentes, por exemplo, já sugerem o uso de clusters de computadores para o processamento e análise de dados de *blockchain* para estabelecer plataformas de gerenciamento de risco voltadas à detecção de atividades ilícitasde maneira mais rápida e eficaz (Zhou et al., 2020; Maheshwari et al., 2023).

Com isso surge a seguinte questão de pesquisa: é possível estabelecer um fluxo de processamento e análise otimizados de dados de *blockchain* que permita identificarendereços de origem de transações ilícitas?

No intuito de endereçar essa questão de pesquisa, este trabalho se baseia em uma revisão sistemática da literatura e procedimentos de simulação computacional para sugerir um fluxo de processamento e análise de volumes massivos de dados do *blockchain* do *bitcoin* com o intuito de avaliar a similaridade de padrões de transações entre dois endereços suspeitos de transações ilícitas.

Este artigo está dividido em cinco seções. A primeira apresenta a contextualização, problemática e objetivos do estudo. A segunda seção contém uma revisãodo estado da arte da pesquisa relacionada com o processamento e análise de dados de *blockchain* e define os

principais conceitos explorados nesta pesquisa. A terceira seção apresenta a metodologia de pesquisa, seguida pelos principais resultados na seção quatro. Por fim, a seção cinco apresenta as conclusões do estudo.

2. Revisão de literatura

A revisão da literatura seguiu uma abordagem de revisão sistemática cujos procedimentos principais estão descritos a seguir.

2.1 Revisão sistemática

O processo de revisão sistemática de literatura teve início com a definição dos conceitos do tópico de pesquisa. Para essa finalidade, as seguintes definições conceituais foram estabelecidas:

- *Blockchain*: redes descentralizadas nas quais seus membros têm total acesso para monitorar todas as transações na rede usando uma abordagem *peer-to-peer*, sendo que tais redes possuem como fundamentos os princípios da imutabilidade, descentralização, segurança e consenso.
- *Big data*: dados caracterizados por grande volume, alta velocidade de geração e grande variedade de formatos.
- Arquitetura de *big data*: tecnologias que demandam métodos de escalabilidade horizontal para processamento eficiente de *big data*.

As bases de busca selecionadas foram a *Scopus* e a *ISI Web of Science*. O motivo da escolha se dá pelo fato delas congregarem boa parte dos periódicos com alto poder de impacto na comunidade científica internacional e, consequentemente, contribuir para um mapeamento mais robusto do estado da arte dos estudos relacionados ao tema de pesquisa.
Uma vez definidos os conceitos do tópico de pesquisa e as bases de busca, foi necessário estabelecer os termos de busca para cada conceito e estruturar as respectivas consultas em cada base. Os termos de busca para cada conceito foram, respectivamente:

- *Blockchain*: *blockchain** OR *bitcoin** OR *crypto*
- *Big data*: *big data* OR *large data*
- Arquitetura de *big data*: *architectur** OR *high performance computing* OR *parallel process** OR *distributed comput** OR *distributed data*

Os textos das consultas utilizadas em cada base estão disponíveis no Apêndice A para facilitar a sua compreensão e reutilização.
Dessa primeira busca foram encontradas 1426 publicações dos mais diferenciados formatos e áreas do conhecimento. A fim de promover a convergência das publicações selecionadas com o objetivo da revisão sistemática de literatura, optou-se por selecionar apenas artigos de periódicos e congressos, os quais, além de representarem as publicações mais recentes da área, também tendem a ser submetidos a um criterioso processo de revisão duplo-cego antes de serem publicados. Além disso, optou-se também por selecionar apenas artigos provenientes das áreas de computação e engenharia, os quais tendem a concentrar discussões mais alinhadas com a questão de pesquisa do presente estudo. Como resultado desses critérios de inclusão, restaram 699 artigos.
Em seguida, teve início um processo mais granular de análise de conteúdo dos artigos, o qual se baseou na exclusão de artigos cujos títulos e resumos revelassem que o foco do estudo não fosse a análise de aspectos das tecnologias de *big data* para processamento e análise de dados de *blockchain*. Também foram excluídos da análise artigos cujo texto integral não estivesse disponível por meio dos acessos institucionais fornecidos às bases

Scopus e *ISI Web of Science*, bem como eventuais artigos que aparecessem em duplicidade nos resultados das buscas realizadas nas duas bases. Nesse último caso, foi mantido apenas o artigo proveniente de uma das bases. Após esse processo de exclusão, restaram 80 artigos.
Como se tratava de um número ainda considerado elevado para uma análise mais detalhada, e a fim de contribuir para maior contemporaneidade do estudo, optou-se também por excluir artigos que tivessem sido publicados antes do período pré-pandemia COVID-19. Essa decisão se baseou no fato de que a intensidade e forma de uso tanto das tecnologias *blockchain* passaram por significativa transformação ao longo desse período (Nguyen, Mai, Bezbradica& Crane, 2022). Consequentemente, para análise final foram considerados 44 artigos de diferentes periódicos e áreas do conhecimento.
O diagrama da Figura 1 ilustra o processo de seleção dos artigos. A partir do método bola-de-neve, essa amostra inicial de artigos foi expandida para suas referências e o conteúdo total utilizado para estabelecer um panorama geral do estado da arte com relação aos tópicos de detecção de atividades ilícitas e fraude envolvendo criptomoedas, desafios no processamento e análise de dados de *blockchain* e arquiteturas de *big data*.

Figura 1 - Diagrama de seleção dos artigos para desenvolvimento da revisão de literatura (Elaborado pelos autores).

2.2 Detecção de atividades ilícitas e fraude envolvendo criptomoedas

A detecção de atividades ilícitas e fraudes é uma das partes essenciais de um sistema de gerenciamento de risco financeiro e tem sido pesquisada há bastante tempo. A fim de identificar tais atividades e minimizar perdas potenciais, os pesquisadores e especialistas do setor se dedicam a descobrir e melhorar os métodos de detecção de atividades ilícitas e fraudes para evitar ao máximo a ocorrência de tais negócios (Zhou et al., 2020).
Tal desafio se faz presente especialmente em operações financeiras envolvendo criptomoedas. Pesquisas recentes sugerem que ao mesmo tempo em que o uso de criptomoedas tem se tornado mais popular, também tem aumentado o seu uso para fins

ilícitos, tais como comércio ilegal de armas, drogas, lavagem de dinheiro e roubos (Maheshwari et al., 2023; Sharma, Agrawal, Bhatia&Tiwari, 2022; Wu et al., 2020). Grande parte desse movimento em direção a transações ilícitas deve-se ao suposto anonimato proporcionado pela tecnologia *blockchain*, na qual os usuários de criptomoedas não precisam se identificar com o uso de informações pessoais.

Contudo, mesmo não fornecendo seus dados pessoais, os usuários de criptomoedas não estão completamente anônimos no*blockchain*. Estudos recentes sugerem que já é possível extrair identificação baseada em comportamento a partir das transações contidas no*blockchain*. O trabalho de Tuner e Irvin (2018), por exemplo, sugere que a análise cuidadosa das transações de criptomoedas com base no uso repetido de chaves públicas específicas associadas a pagamentos pode ser utilizada para mapear as transações de um determinado usuário na rede. Tais transações, por sua vez, podem ser usadas para estabelecer os padrões de compra e venda único de um determinado usuário, os quais podem ser utilizados subsequentemente para identificar esse usuário em um futuro conjunto de transações na rede.

Mais recentemente, o trabalho de Maheshwari et al. (2023) sugeriu que biometria comportamental pode ser útil para associar os endereços de origem das transaçõesde criptomoedas com endereços fraudulentos previamente conhecidos. A ideia se baseia no princípio de que o intervalo de tempo aleatório ou RTI (do termo em inglês *Random Time Interval*) entre transações decriptomoedas realizadas por um fraudador poderia ser utilizado como uma espécie de impressão digital que permita identificar outros endereços de transações eventualmente associados a esse indivíduo. O conceito de RTI foi utilizado pela primeira vez no artigo de Laskaris, Zafeiriou e Garefa(2009), no qual os participantes do estudo receberam um botão para pressionar aleatoriamente e foi demonstrado que a série temporal produzida para cada indivíduo poderia ser usada como uma assinatura biométrica do processo de pensamento cognitivo de uma pessoa. No caso dascriptomoedas, à medida que o fraudador realiza transações ilícitas ou ilegais com um endereço específico, as informações de data e hora das transações ficam registradas e disponíveis publicamente no*blockchain*e podem ser usados para análise de RTI com outros endereços previamente conhecidos e associados com transações ilícitas ou fraudulentas.

No entanto, o rápido desenvolvimento das tecnologias *blockchain*, somado ao seu volume de dados e ao crescente uso das criptomoedas torna a execução de análises envolvendo dados contidos no*blockchain* uma atividade extremamente desafiadora do ponto de vista computacional. Essa falta de capacidade em processar e analisar dados do *blockchain*, por sua vez, tendem a limitar a capacidade das organizações em desenvolver soluções de detecção de atividades ilícitas ou antifraude que atendam satisfatoriamente às demandas atuais do mercado (Zhou et al., 2020).

2.3 O*bitcoin* e o processamento de dados de *blockchain*

O*bitcoin* tem seu início datado em 2009 e representa a implementação de criptomoeda dominante (Nakamoto, 2008). O sistema se apoia em um protocolo acordado para a transmissão de trocas de valor entre participantes de uma rede *peer-to-peer*. Esses registros de transações são verificados regularmente por nós especializados em mineração na rede e cuja confiabilidade é garantida por meio do *blockchain*, um banco de dados à prova de adulteração distribuído publicamente. Tal banco de dados e suas atualizações são públicos para permitir a formação de um consenso majoritário em tempo real quanto ao estado atual do sistema válido. Desta forma, através da combinação de criptografia com incentivos econômicos, participantes associados a pseudônimos são capazes de estabelecer confiança

mútua e conduzir transações seguras entre si (McGinn et al., 2016; McGinn, McIlwraith&Guo, 2018; Maheshwari et al., 2023).

Todo o sistema *bitcoin* existe exclusivamente para criar, propagar, verificar e registrar estruturas de dados conhecidas como transações. Uma transação é um registro atômico por meio do qual a propriedade de uma quantidade de *bitcoin* é transferida pelo proprietário atual para um novo proprietário. Uma transação é composta de 1.n saídas e 0.n entradas. As saídas da transação são novos registros de quantidades de *bitcoin* associados a um determinado endereço *bitcoin*, sendo uma representação do componente de chave pública de um desafio criptográfico assimétrico que pode ser satisfeito apenas pelo novo proprietário. As entradas de transação são ponteiros para saídas de transação não gastas (sigla em inglês UTXO's) existentes, juntamente com uma prova válida do desafio criptográfico existente do UTXO específico para demonstrar a propriedade de forma verificável. É somente através do fornecimento de todas as soluções de entrada para os desafios criptográficos que uma transação será reconhecida e registrada pelos participantes como válida, evitando roubos. Da mesma forma, qualquer transação que tente reatribuir a propriedade de valores anteriormente não onerados (gasto duplo) ou incluir saídas somando mais do que as entradas (falsificação) será rejeitada pela maioria dos participantes honestos. As saídas não gastas de cada nova transação podem, portanto, ser consideradas a fronteira de uma árvore específica de gastos em todo o gráfico de transações, enraizado em um conjunto de transações de base de moeda. As transações são transmitidas pela rede e cada nó participante mantém uma cópia das transações recebidas que considera válidas em uma estrutura de dados mantida em memória volátil (McGinn et al., 2016). Além disso, cada usuário de *bitcoin* tem acesso a uma chave privada e uma chave pública. O endereço *bitcoin* é derivado da chave pública pelo uso de *hashing* criptográfico unidirecional e pode representar a origem e o destino do *bitcoin* nas transações (Antonopoulos, 2017).

O *blockchain*, por sua vez, é armazenado em arquivos binários blk.dat localizados nos nós computacionais pertencentes à rede de *bitcoin*. Cada arquivo blk.dat possui um tamanho de 128 MB e contém blocos de dados brutos que são recebidos pelo nó da rede de *bitcoin*. Os dados contidos nos arquivos blk.dat estão armazenados em formato binário, conforme ilustrado na Figura 2.


```
F9 BE B4 D9 1D 01 00 00 01 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
3B A3 ED FD 7A 7B 12 B2 7A C7 2C 3E 67 76 8F 61 7F C8 1B C3 88 8A 51 32 3A 9F B8 AA 4B 1E 5E 4A 29 AB 5F 49 FF FF 00 1D 1D AC 2B 7C
01 01 00 00 00 01 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 FF FF FF FF 4D 04
FF FF 00 1D 01 04 45 54 68 65 20 54 69 6D 65 73 20 30 33 2F 4A 61 6E 2F 32 30 30 39 20 43 68 61 6E 63 65 6C 6C 6F 72 20 6F 6E 20 62
72 69 6E 6B 20 6F 66 20 73 65 63 6F 6E 64 20 62 61 69 6C 6F 75 74 20 66 6F 72 20 62 61 6E 6B 73 FF FF FF FF 01 00 F2 05 2A 01 00 00
00 43 41 04 67 8A FD B0 FE 55 48 27 19 67 F1 A6 71 30 B7 10 5C D6 A8 28 E0 39 09 A6 79 62 E0 EA 1F 61 DE B6 49 F6 BC 3F 4C EF 38 C4
F3 55 04 E5 1E C1 12 DE 5C 38 4D F7 BA 0B 8D 57 8A 4C 70 2B 6B F1 1D 5F AC 00 00 00 00
```


Figura 2 - Os primeiros 293 *bytes* do arquivo blk0000.dat (Elaborado pelos autores).

Cada bloco constituinte do arquivo blk.dat possui uma estrutura binária que pode ser dividida em cinco partes principais:

- *Magic bytes*: constituídos pela sequência inicial de quatro *bytes* do bloco edetermina a posição inicial e final de cada bloco.
- *Size*: representado pela sequência de quatro *bytes*após os *magic bytes*e especifica o tamanho do bloco
- *Block header*: representado pela sequência de 80 *bytes* após o *size* e contém metadados diversos, tais como versão do bloco (4 bytes), *hash* do bloco anterior (32 bytes), *hash* de todas as transações do bloco (32 bytes), hora em que o bloco foi minerado (4 bytes), o alvo da mineração (4 bytes) e o campo para mineração (4 bytes).

- *Txcount*: sequência de *bytes* de tamanho variável após o *block header*e representa o número de transações existentes no bloco
- *Transaction data*: sequência de *bytes* após o *txcount* e contém o *hash* das informações das transações.

Os desafios relacionados ao processamento e análise de dados contidos em *blockchain* são reportados na literatura desde meados de 2015 quando o*blockchain*do *bitcoin* possuía um tamanho de, aproximadamente, 50 GB (Rubin, 2015). Àquela época, mesmo representando um volume de dados viável de ser manipulado em memória, a sua expansão de formato binário para uma estrutura de dados relacional que permitisse uma análise mais detalhada das transações já demandava recursos computacionais consideráveis.

Dados recentes de meados de 2022, por outro lado, já sugerem que o tamanho atual do *blockchain*do *bitcoin* está próximo de 400 GB (Maheshwari et al., 2023), o que torna o problema de detecção de fraude, na verdade, um problema de *big data*.

O trabalho de Zhou et al. (2020) ilustra bem esse desafio. Objetivando detectar fraudes financeiras em uma base de dados de *blockchain* e usando linguagem de programação R em apenas um computador, o tempo de computação foi superior a 60 horas. Por outro lado, o mesmo trabalho sugere que abordagens de processamento paralelo e dados distribuídos podem acelerar significativamente esse processo.

2.4 Arquiteturas de *big data*

Big data pode ser definido como um novo paradigma tecnológico impulsionado por dados com alto volume, velocidade, variedade e veracidade. Tais dados advêm de várias fontes que incluem a Internet, dispositivos móveis, mídias sociais, dispositivos geoespaciais, sensores e outros dados gerados por máquinas. Extrair valor do*big data* permite que os negócios percebam e respondam melhor ao ambiente econômico e esta capacidade tem se tornando um elemento chave para criar vantagens competitivas em um mercado complexo e em rápida mudança (Chan, 2013).

Do ponto de vista tecnológico, o armazenamento e processamento de dados tradicional, bem como a análise de dados estruturados usando sistemas gerenciadores de bancos de dados relacionais (RDBMS) não satisfazem mais os desafios do*big data*. As tendências de tecnologia para *big data* abrangem *software* de código aberto, servidores *commodities* e plataformas de processamento massivamente distribuídas em paralelo (Chan, 2013; Zhou et al., 2020; Maheshwari et al., 2023).

As pesquisas mais recentes sugerem que o uso de plataformas de *big data*, amplamente apoiadas em paradigmas de computação paralela e distribuída parecem constituir uma solução atrativa para o processamento e análise de dados de *blockchain*. O uso do *Apache Spark*, por exemplo, foi reportado no trabalho de Zhou et al. (2020) para análise massiva de dados de*blockchain*naidentificação de fraude financeira em uma cadeia de suprimentos. Em primeiro lugar, os dados do *blockchain* foram divididos e armazenados na forma de arquivos HDFS (do termo em inglês, *HadoopDistributed FileSystem*) nos nós do cluster, os quais são disponibilizados para manipulação por meio do conjunto de dados distribuído resiliente *Spark*ou RDD (do termo em inglês *ResilientDistributedDataset*). Nessa arquitetura de sistema, o tempo de computação para análises de fraudes financeiras pode chegar a 1 hora.

Já o uso de linguagens de programação mais eficientes, tais como o C++ para o processamento de dados brutos do *blockchain*do *bitcoin* encontra suporte no trabalho de McGinn, McIlwraith e Guo (2018). Nesse estudo, os autores desenvolveram um analisador personalizado para processar e estruturar rapidamente os arquivos binários brutos em um

cluster HPC (do termo em inglês *High Performance Computing*) de 400 nós. Com essa arquitetura, à época do estudo, foi possível extrair 8 anos de dados de transações de maneira eficiente, uma vez que os dados do *blockchain* naturalmente apresentam uma característica de paralelismo: cada arquivo de dados binários bruto de 128 MB contém blocos e transações que são únicos e estão relacionados entre si por meio de outros identificadores únicos.
Considerados em conjunto, os estudos encontrados até o momento sugerem que a combinação de linguagens de programação tradicionalmente mais eficientes em termos de processamento computacional, tais como o C++, associados ao uso de processamento paralelo e distribuído dos dados, parecem ser uma alternativa viável para o desenvolvimento de uma arquitetura de *big data* otimizada para o processamento e análise de dados brutos do *blockchain*do *bitcoin*.

3. Metodologia de pesquisa

Para atingir os objetivos desteestudo, optou-se por uma abordagem de simulação, ou seja, o uso de técnicas computacionais para simular o funcionamento de sistemas produtivos a partir de modelos simplificados. Nesse sentido, o processo de simulação geralmente envolve o desenvolvimento de um modelo que represente de maneira fiel a realidade a ser estudada. Para esse fim, podem ser utilizados ambientes computacionais com o auxílio de *software*s disponíveis no mercado para execução de testes e geração de amostras representativas do cenário sendo simulado. A partir da análise de tais amostras, o pesquisador pode então avaliar o desempenho do sistema em estudo(Gerhardt& Silveira, 2009).
No presente estudo foi elaborado um ambiente de simulação usando recursos de *hardware* em nuvem pública e recursos de *software*de código aberto para estabelecimento de uma plataforma de processamento e análise de volumes massivos de dados. As escolhas dos recursos de *hardware* e *software* utilizados no ambiente de simulação foram pautadas na revisão de literatura e se basearam no estado da arte dos estudos envolvendo volumes massivos de dados.
O ambiente de simulação se apoia na proposta de Maheshwari et al. (2023) e busca estabelecer um fluxo de extração, processamento e análise de dados do *blockchain*do *bitcoin* que permitam estabelecer uma análise de similaridade da RTI entre um endereço qualquer do *blockchain* e endereços previamente associados a atividades ilícitas. Para essa finalidade, é fundamental não só extrair o RTI dos dados brutos do *blockchain* do*bitcoin* mas também possuir o RTI de endereços associados a atividades ilícitas.
A descrição detalhada do ambiente e os recursos utilizados é apresentada na seção a seguir e os códigos desenvolvidos estão disponíveis em um repositório do *GitHub*: https://github.com/HWatanuki/Trabalho_D2TEC_Final

4. Resultados

O processo de extração e transformação dos dados brutos do *blockchain* do *bitcoin* envolveu o uso da plataforma *HPCC Systems*. Trata-se de uma plataforma de *big data* de código aberto que adota um paradigma de paralelismo de fluxo de dados voltado para a otimização da manipulação de volumes massivos de dados (Xu, Muharemagic, Villanustre&Apon, 2017). Além do processamento paralelo e distribuído, uma das grandes vantagens oferecidas pela plataforma é o uso de uma linguagem de manipulação de dados denominada ECL (do termo em inglês, *Enterprise ControlLanguage*)que compila para

C++, o que tende a contribuir para uma execução mais eficiente de códigos que envolvam a manipulação massiva de dados, como no caso do *blockchain* do *bitcoin*.
A plataforma foi configurada em uma infraestrutura de nuvem *Microsoft Azure*que continhaos seguintes recursos principais (Figura 3):

- Uma rede virtual (vnet) padrão da *Azure* com *subnets* pública e privada.
- Uma conta de armazenamento de propósito geral V2 com redundância para armazenamento de dados e acessível por meio de *Azure Files Shares*.
- Um *cluster*auto escalável *Kubernetes*gerenciado pelo *Azure Kubernetes Service* (AKS) e constituído porduas instâncias Standard_B2s com 2 vCPUs e 4 GB de memória para processamento dos dados e uma instância Standard_D2_v4 com 2 vCPUs e 8 GB de memória para gerenciamento do *cluster*.

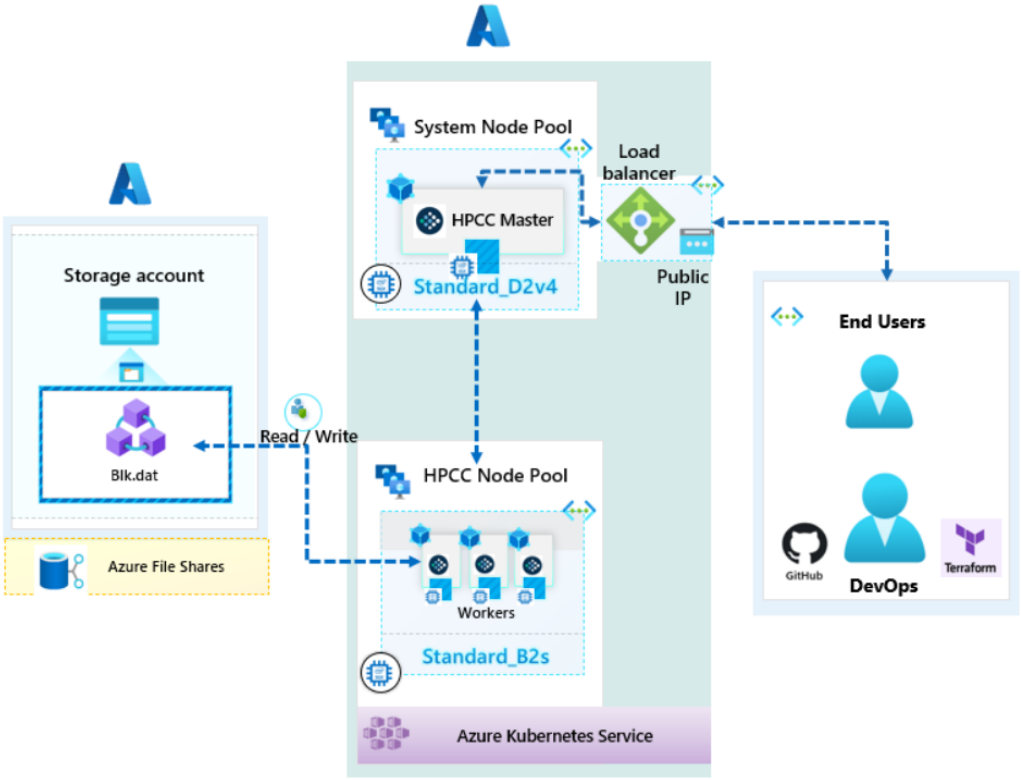

Figura 3 - Diagrama de arquitetura do ambiente de simulação utilizado (Elaborado pelos autores).

O fluxo completo de processamento e análise dos dados brutos do *blockchain* do *bitcoin* seguiu um processo de quatro etapas, conforme descrito a seguir.

4.1 Extração dos dados brutos do *blockchain*

Uma instância foi configurada usando o *softwareBitcoin Core*(Nakamoto, 2008) para fazer o download dos dados brutos do *blockchain* do *bitcoin*. O *blockchain* foi dividido em arquivos blk.dat de 128 MB e importados na plataforma HPCC Systems como *blob's* (do termo em inglês, *binarylargeobject*).
Em seguida, a fim de estruturar os dados binários de transações *bitcoin* presentes nos blocos dos arquivos blk.dat, um analisador foi desenvolvido para converter os dados binários em arquivos semiestruturados em formato CSV (do termo em inglês, *commaseparatedvalues*).

Com base no processamento dos blocos de transações do *blockchain* do *bitcoin*, o analisador foi capaz de extrair e estruturaros seguintes dados das transações:

1. *tx_hash*: *hash* da transação incluída no bloco.
2. *in_index*: índice de saída da transação anterior que originou a entrada da transação atual.
3. *in_hash*: *hash*da transação que originou a entrada da transação atual.
4. *out_index*: índice da saída da transação.
5. *out_addr*: endereço de destino da transação.
6. *out_val*: valor em *bitcoin* negociado.
7. *timestamp*: data e hora do bloco.

A figura 4 apresenta uma ilustração da saída dos dados das transações estruturados em formato relacional.

tx_hash	in_index	in_hash	out_index	out_addr	out_val	timestamp
4a5e1e4baab89f3a325…	4294967295	000000000000000…	0	1A1zP1eP5QGefi2DMPTfTL5…	5000000000	2009-01-03 18:15:05
0e3e2357e806b6cdb1f…	4294967295	000000000000000…	0	12c6DSiU4Rq3P4ZxziKxzrL…	5000000000	2009-01-09 02:54:25
9b0fc92260312ce44e7…	4294967295	000000000000000…	0	1HLoD9E4SDFFPDiYfNYnkBL…	5000000000	2009-01-09 02:55:44
999e1c837c76a1b7fbb…	4294967295	000000000000000…	0	1FvzCLoTPGANNjWoUo6jUGu…	5000000000	2009-01-09 03:02:53
df2b060fa2e5e9c8ed5…	4294967295	000000000000000…	0	15ubicBBWFnvoZLT7GiU2qx…	5000000000	2009-01-09 03:16:28
63522845d294ee9b018…	4294967295	000000000000000…	0	1JfbZRwdDHKZmuiZgYArJZh…	5000000000	2009-01-09 03:23:48
20251a76e64e920e582…	4294967295	000000000000000…	0	1GkQmKAmHtNfnD3LHhTkewJ…	5000000000	2009-01-09 03:29:49
8aa673bc752f2851fd6…	4294967295	000000000000000…	0	16LoW7y83wtawMg5XmT4M3Q…	5000000000	2009-01-09 03:39:29
a6f7f1c0dad0f2eb6b1…	4294967295	000000000000000…	0	1J6PYEzr4CUoGbnXrELyHsz…	5000000000	2009-01-09 03:45:43
0437cd7f8525ceed232…	4294967295	000000000000000…	0	12cbQLTFMXRnSzktFkuoG3e…	5000000000	2009-01-09 03:54:39

Figura 4 - Representação de 10 registros da estrutura de dados gerada pelo analisador do *blockchain* do *bitcoin* (Elaborado pelos autores).

4.2Transformação dos dados estruturados

Como o cálculo do RTI depende da informação do endereço de origem das transações *bitcoin* e esse dado não está prontamente disponível nos blocos de transação do *blockchain* do *bitcoin*, foi necessário realizar uma transformação nos dados extraídos na etapa anterior. A lógica da transformação consistiu em utilizar os campos de “in_index” e “in_*hash*” de cada transação, respectivamente, o índice de saída da transação anterior que originou a entrada da transação atual e o *hash* da transação anterior que originou a entrada da transação atual, para associar as transações contidas nos blocos com suas transações anteriores, uma vez que o endereço de destino da transação anterior representa o endereço de entrada da transação contida no bloco. Esse processo é ilustrado nafigura 5 e permite que os endereços de entrada de cada transação contida no bloco sejam determinados.

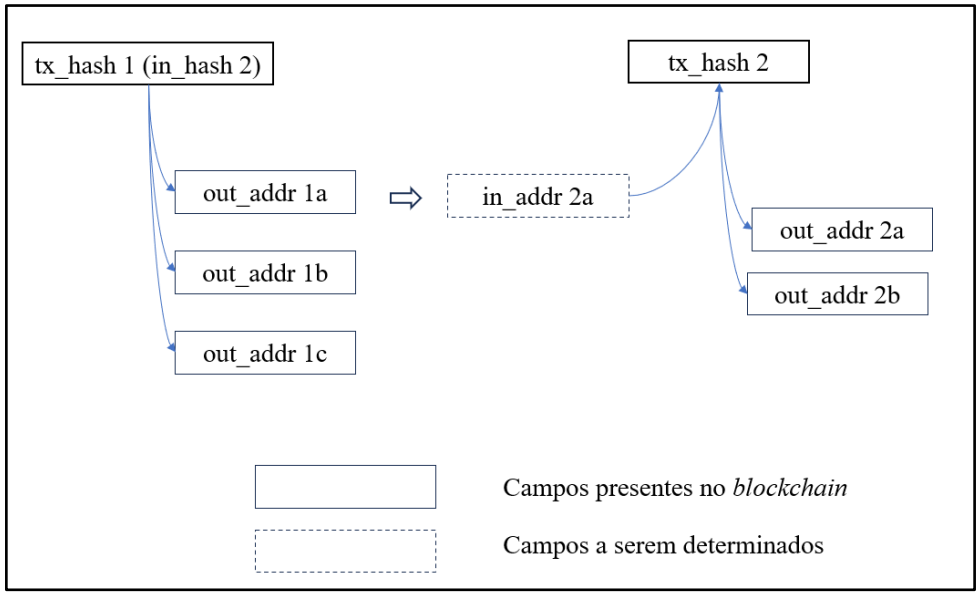

Figura 5 - Relação entre endereços de entrada e saída das transações contidas na *blockchain* do *bitcoin* (Elaborado pelos autores).

A figura 6 ilustra a estruturação dos dados obtida, na qual cada registro do conjunto de dados transformado agora representa uma transação de *bitcoin* com seus respectivos endereços de entrada e saída.

tx_hash	in_addr	out_addr	out_val	unix_time	timestamp
7f24cbde4ac486e8f72b365…	11122zFBGFXUzFSjx5…	1QB5eMV2FzbgyZBAvumFnJ7WKW…	32000	1343058682	2012-07-23 15:51:22
7f24cbde4ac486e8f72b365…	11122zFBGFXUzFSjx5…	1DV3taToMqbQHkdpey4Ww5iB6a…	1529134000	1343058682	2012-07-23 15:51:22
a443aa87ea1d8c209f66d07…	11123WHq5dD3rFP8f4…	135u2m38X6atesHsJApu3vRGiB…	7770000	1334056358	2012-04-10 11:12:38
a443aa87ea1d8c209f66d07…	11123WHq5dD3rFP8f4…	1Gi8afmdDcSWcnfMeHZ3r6ieGc…	1999950000	1334056358	2012-04-10 11:12:38
2642af65550972b0a7b3333…	11129QtW1GqAQon2z8…	1GMRwymmE5nqbwdr2ue5Mwc76X…	1000000	1340915303	2012-06-28 20:28:23
2642af65550972b0a7b3333…	11129QtW1GqAQon2z8…	1QJaZWM9FVvSiBaVbWGzAgpakw…	1566000000	1340915303	2012-06-28 20:28:23
2642af65550972b0a7b3333…	11129QtW1GqAQon2z8…	1GMRwymmE5nqbwdr2ue5Mwc76X…	1000000	1340915303	2012-06-28 20:28:23
2642af65550972b0a7b3333…	11129QtW1GqAQon2z8…	1QJaZWM9FVvSiBaVbWGzAgpakw…	1566000000	1340915303	2012-06-28 20:28:23
57d92a6f91844d6a58e51a8…	1114DgPd62D8RPZ1Nb…	169v7QZDW4bS3nFbrijQksVKbb…	4329697673	1313928513	2011-08-21 12:08:33
57d92a6f91844d6a58e51a8…	1114DgPd62D8RPZ1Nb…	15xmacS6RpVbVE76HmpAnPkHPG…	439463902	1313928513	2011-08-21 12:08:33

Figura 6–Representação de 10 registros da estrutura transformada do conjunto de transações do *blockchain* do *bitcoin* (Elaborado pelos autores).

4.3 Geração do RTI

Uma vez identificados os endereços de origem de cada transação contida no *blockchain* do *bitcoin*, foi necessário estabelecer o RTI para cada endereço. Para essa finalidade, foram utilizados os registros de data e hora das transações e o RTI para cada endereço foi calculado considerando-se a diferença entre a data e hora das sequências de transações feitas a partir daquele endereço. Vale ressaltar que a *blockchain* do *bitcoin*registra apenas a data e hora do bloco de transações e não de cada transação específica e por issofoi necessário aproximar a data e hora da transação com a data e hora da geração do bloco, o que tende a ser uma aproximação realista para análises dessa natureza (Maheshwari et al., 2023).A figura 7 ilustra a estruturação dos dados obtida, na qual cada registro do conjunto de dados transformado agora representa um endereço de origem de uma transação de *bitcoin* com seu respectivo RTI.

in_addr	rti
	Item
111213xkVmeHT1v7N7Ff43XmVfLAUq7WEZ	0,745477,0,0,0,1073155,0,2081033,464738,0,0,118044,0,0,0,0,0,196561,0,0,0,0,0,0,0,295522,0,0,0,0,0,0,0,0,0,…
111mamKazi1ofZ854rKTcAxhTVuGDPhhL	0,0,345680,0,0,0,93923,0,85835,0,0,0,0,0,0,0,70206,0,0,0,0,0,5603,0,82078,0,0,0,16954,0,59462,0,0,18491,0,0…
112345FMq98vjXZTbet2nYHtvqX8p9qQKt	0,0,1931,0,0,0,0,0,33009,0,0,0,0,0,46116,0,9312,0,0,0,0,0,665,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,894,0,0,0,0,0,4…
112KFspUbaqEadpkBa6W5741xP5oFFGo16	0,0,3865232,0,4660,0,8189,0,3786,0,55535,0,5284,0,0,0,26926,0,141417,0,124101,0,62686,0,6971,0,511692,0,201…
112P8dAd84bzYhVopnTU1tVCwbAcUptkck	0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0…
112PhXdxVAXXZPA9aYhvZNNdbLUvZ4rbSH	0,0,0,0,0,0,0,0,0,0,0,0,46925,0,0,0,0,0,0,0,0,0,0,0,0,0,913377,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,79…
112Pm3cd2LgS7GJT4RaiqnybKHKzCTtGY7	0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0…
112QNTtSrKgqua5s7uCDHasJKqXsaHg93R	0,0,147866,0,192168,0,323355,0,60734,0,11238,0,9266,0,34278,0,9145,0,0,0,14676,0,4476,0,25521,0,198,0,12525…
112RrU2ZjhYGjwAwhgFdg6Q9dSHsZwhMko	0,0,0,0,249702,0,0,0,256118,0,0,0,0,0,0,0,341891,0,0,0,0,0,9813,0,82049,0,172212,0,0,0,428592,0,109007,0,0,…
112SpuCBQijmr3hxJQR5ZLEUjv1BcbHQwm	0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,253003,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,…

Figura 7 –Representação de 10 registros da estrutura de dados contendo os endereços de origem da transação e seus respectivosRTI’s (Elaborado pelos autores).

Esse método de obtenção do RTI foi utilizado tanto para os endereços do *blockchain* do *bitcoin*, como também para os endereços previamente associados a comportamentos ilícitos ou fraudes envolvendo *bitcoin*.

Para obter os endereços de referência ilegais foi utilizado o conjunto de dados conhecido como BABD (do termo em inglês *BitcoinAddressBehaviorDataset*) referenciado no estudo de Xiang et al. (2022). Nesse conjunto de dados, os endereços legais são endereços que foram obtidos aleatoriamente com a suposição de que 1% dos endereços *bitcoin*estão envolvidos em atividades ilícitas ou fraudes. Esses últimos são divididos em 13 categorias, cada uma contendo uma lista de endereços de um tipo diferente de crime. Os endereços ilegais de referência foram escolhidos a partir deles de maneira aleatória.

4.4 Comparação de similaridade dos endereços

Após a obtenção do RTI tanto do endereço a ser avaliado como dos endereços envolvidos em atividades ilícitas, um teste estatístico pode ser aplicado para se verificar se os conjuntos de observações foram extraídos da mesma distribuição. Para isso, foi escolhido o teste de Kolmogorov-Smirnov (Hodges, 1958), o qualutiliza uma variação do teste de uma amostra que compara um conjunto de duas observações e cria uma distribuição cumulativa para ambos os conjuntos de observações após classificá-los. A diferença entre as duas distribuições é obtida e o máximo destas diferenças é comparado com o valor crítico. Caso a estatística do teste de Kolmogorov-Smirnov seja inferior ao valor crítico, é rejeitada a hipótese nula de que ambas as observações se originam da mesma distribuição.

No presente estudo, amostras aleatórias com RTI de66 endereços de origem de transações de *bitcoin* associados a atividades ilícitas foram comparadas com o RTI de 263 endereços selecionados da base BABD por meio do teste de Kolmogorov-Smirnov.Com base nesse teste foi possível obter uma acuracidade de 0,85 (*f1-score* de 0,92), o que representa um resultado inicial promissor para a abordagem e sugere que usuários envolvidos em atividades ilegais podem ser detectados pelo uso de endereços ilegais previamente conhecidos.

5. Considerações finais

Nos últimos anos, um volume crescente de transações envolvendo criptomoedas têm sido utilizadas em diversas atividades ilegais, incluindo comércio ilegal de drogas earmas, roubos, fraudes e outros crimes graves. De acordo com umrelatório recente (Grauer, Kueshner&Updegrave, 2022), um volume total de US$ 14 bilhões em criptomoedas esteve envolvido em atividades ilícitas no ano de 2021, mas um volume provavelmente maiorpoderia ser identificado caso abordagens mais avançadas e precisas permitissem monitorar os endereços de origem das atividades ilícitas.

Este artigo apresenta umaabordagem para processar e analisar volumes massivos de dados do *blockchain* do *bitcoin* com o intuito de avaliar a similaridade de padrões de transações entre dois endereços suspeitos de transações ilícitas. Para essa finalidade, foi desenvolvido um ambiente de simulação computacional que permite extrair e transformar os dados binários do *blockchain* do *bitcoin* e estruturá-los para que o RTI dos endereços de origem das transações sejam obtidos. O RTI de tais endereços podem então ser comparados com o RTI de endereços previamente associados a atividades ilícitas com o intuito de identificar a carteira de endereços relacionadas com indivíduos que praticam atividades ilícitas usando criptomoedas.

Embora represente um esforço de pesquisa inicial nesse contexto, espera-se que este trabalho contribua comorganizações que enfrentam a necessidade crescente do uso de dados de *blockchain* para suportar o desenvolvimento de soluções antifraude compatíveis com as necessidades do mercado da atualidade.Essa potencial contribuição possui dimensões sociais, científicas e tecnológicas. Do ponto de vista social, espera-se que a análise eficiente de dados de *blockchain* permita um combate mais efetivo não só das operações de fraudes financeiras, mas também de outros comportamentos ilícitos que são fomentados pelo relativo anonimato dos usuários de criptomoedas. Do ponto de vista científico, espera-se que o estudo contribua com os esforços de pesquisa na área de computação e engenharia atualmente dependentes da identificação de soluções que permitam o processamento e análise otimizados de dados de *blockchain*(Rubin, 2015; McGinn, McIlwraith&Guo, 2018; Zhou et al., 2020;Maheshwari et al., 2023). Por fim, da perspectiva tecnológica, espera-se que o estudo contribua com a identificação e desenvolvimento da combinação de recursos de *hardware* e *software* mais otimizada para o endereçamento de um problema puramente de natureza computacional envolvendo volume massivo de dados.

Por fim, é importante destacar as limitações do estudo, as quais também podem proporcionar oportunidades de estudos futuros. Em primeiro lugar, é importante salientar que a abordagem desenvolvida no presente estudo depende primariamente de um conjunto de transações e endereços de origem associados a atividades ilícitas, o que nem sempre está prontamente disponível. O conjunto de dados BABD utilizado neste estudo fornece alguns endereços ilegais, mas apenas com um grau limitado de certezae estima-se que níveis maiores de acuracidade poderiam ser almejados com base em dados mais precisos ou que considerem outros atributos das transações além apenas do RTI, como por exemplo, os valores transacionados. Em segundo lugar, para contribuir com um melhor entendimento da potencialidade da abordagem apresentada nesse estudo, sugere-se a realização de testes adicionais que envolvam não só amostras com endereços associados a atividades ilícitas, mas tambémendereços que não estejam associados a tais atividades, assim como a eventual utilização deoutros critérios de ajuste. Por fim,embora o ambiente de simulação utilizado no estudo tenha apresentado resultados promissores, estudos adicionais são sugeridos para avaliar como diferentes parâmetros e configurações de arquitetura de *big data* poderiam otimizar ainda mais os processos aqui apresentados para o processamento e análise do *blockchain* do *bitcoin*.

Referências bibliográficas

Akcora, C. G., Dixon, M. F., Gel, Y. R., &Kantarcioglu, M. (2018). Blockchain data analytics. *Intelligent Informatics*, 4.

Antonopoulos, A. M. (2017). *Mastering Bitcoin: Programming the open blockchain*. California: O'Reilly.

Chan, J. O. (2013). An architecture for big data analytics. *Communications of the IIMA*, 13(2), 1-13.

Deepa, N., Pham, Q. V., Nguyen, D. C., Bhattacharya, S., Prabadevi, B., Gadekallu, T. R., Maddikunta, P. K. R., Fang, F., & Pathirana, P. N. (2022). A survey on blockchain for big data: Approaches, opportunities, and future directions. *Future Generation Computer Systems*, 131, 209-226.

Emery, J. A., &Latapy, M. (2021). Full Bitcoin blockchain data made easy. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 240-243.

Gerhardt, T. E., & Silveira, D. T. (2009). *Métodos de pesquisa*. Plageder.

Grauer, K.,Kueshner, W.,&Updegrave, H. (2022).*The 2022 CryptoCrime Report*.Disponível em:https://go.chainalysis.com/2022-Crypto-Crime-Report.html.

Hodges Jr, J. L. (1958). The significance probability of the Smirnov two-sample test. *Arkivförmatematik*, 3(5), 469-486.

Laskaris, N. A., Zafeiriou, S. P., &Garefa, L. (2009). Use of random time-intervals (RTIs) generation for biometric verification. *Pattern Recognition*, 42(11), 2787-2796.

Maheshwari, R., Shobha, G., Shetty, J., Chala, A., & Watanuki, H. (2023). Illicit Activity Detection in Bitcoin Transactions using Timeseries Analysis. *International Journal of Advanced Computer Science and Applications*, 14(3), 13-18.

McGinn, D., Birch, D., Akroyd, D., Molina-Solana, M., Guo, Y., & Knottenbelt, W. J. (2016). Visualizing dynamic bitcoin transaction patterns. *Big data*, 4(2), 109-119.

McGinn, D., McIlwraith, D., & Guo, Y. (2018). Towards open data blockchain analytics: a Bitcoin perspective. *Royal Society open science*, 5(8), 1802981-14.

Nakamoto, S. (2008).*Bitcoin: A peer-to-peer electronic cash systems*.Disponível em: https://bitcoin.org/bitcoin.pdf

Nguyen, A. P. N., Mai, T. T., Bezbradica, M., & Crane, M. (2022). The cryptocurrency market in transition before and after covid-19: An opportunity for investors? *Entropy*, 24(9), 1317-1345.

Rubin, J. (2015). *Btcspark: Scalable analysis of the bitcoin blockchain using spark*. Disponívelem: https://rubin.io/public/pdfs/s897report.pdf

Sharma, A., Agrawal, A., Bhatia, A., &Tiwari, K. (2022). Bitcoin's blockchain data analytics: A graph theoretic perspective. *Proceedings of the International Conference on Advanced Information Networking and Applications,* 459-470.

Turner, A., & Irwin, A. S. M. (2018). Bitcoin transactions: a digital discovery of illicit activity on the blockchain. *Journal of Financial Crime*, 25(1), 109-130.

Wu, Y., Tao, F., Liu, L., Gu, J., Panneerselvam, J., Zhu, R., & Shahzad, M. N. (2020). A bitcoin transaction network analytic method for future blockchain forensic investigation. *IEEE Transactions on Network Science and Engineering*, 8(2), 1230-1241.

Xiang, Y., Lei, Y., Bao, D., Ren, W., Li, T., Yang, Q., Liu, T., Zhu, T.,& Choo, K. K. R. (2022). *Babd: A bitcoin address behavior dataset for pattern analysis*. arXiv preprint arXiv:2204.05746.

Xu, L., Muharemagic, E., Villanustre, F., & Apon, A. (2017). ECL-watch: a big data application performance tuning tool in the HPCC systems platform. *Proceedings of the IEEE International Conference on Big Data, USA*, 2941-2950.

Zhou, H., Sun, G., Fu, S., Fan, X., Jiang, W., Hu, S., & Li, L. (2020). A distributed approach of big data mining for financial fraud detection in a supply chain. *Computers,Materials& Continua*, 64(2), 1091-1105.

APÊNDICE A – Consultas utilizadas para busca de publicações

A fim de encontrar publicações que tratassem do tema de interesse, as seguintes lógicas de busca foram elaboradas para cada base de publicações, respectivamente:

- Scopus: `(TITLE-ABS-KEY ( blockchain* OR bitcoin* OR crypto ) AND TITLE-ABS-KEY ( big AND data OR large AND data ) AND TITLE-ABS-KEY ( architectur* OR high AND performance AND computing OR parallel AND process* OR distributed AND comput* OR distributed AND data ) )`
- ISI Web of Science:`blockchain* OR bitcoin* OR crypto (Topic) AND big data OR large data (Topic) AND architectur* OR high performance computing OR parallel process* OR distributed comput* OR distributed data (Topic)`