FACULDADE SANTÍSSIMO SACRAMENTO

CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

ALAGOINHAS – BA
2023

BIG DATA NA CIBERSEGURANÇA: Explorando o potencial da análise de Big Data na prevenção de violações de dados causadas por ameaças externas

Ricardo de Souza Rabelo Filho [1]

Joan Marcel Couto de Melo [2]

Gabriela Viana G. de Noronha [3]

RESUMO

O uso da Big Data na cibersegurança é uma solução emergente para tratar do alto volume de dados que soluções tradicionais não conseguem tratar e que pode ajudar as organizações a serem mais eficazes e rápidas na prevenção e resposta na tentativa de violações de dados. Este trabalho tem como proposta explorar o uso da Big Data na cibersegurança, com o objetivo principal de identificar soluções de defesa cibernética baseadas em Big Data que possam substituir o SIEM (Gerenciamento e Correlação de Eventos de Segurança). Para atingir o objetivo da pesquisa, foram reunidos conceitos relevantes relacionados a Big Data, violações de dados e SIEM para melhor entendimento acerca do tema. Foi feito uma revisão sistemática de artigos relacionados ao tema, foram apresentadas soluções e arquiteturas de cibersegurança baseadas em Big Data e foi feita uma análise de estudo de caso existente em uma empresa que oferece serviços de cibersegurança baseadas em Big Data. O objetivo da pesquisa foi alcançado, foram identificadas soluções de segurança cibernética baseadas em Big Data. No entanto, foi observado que a utilização de Big Data não substitui integralmente o SIEM, mas sim representa uma evolução dessa abordagem. Além disso, é relevante ressaltar a existência de soluções comerciais que integram tanto o SIEM quanto técnicas de análise de Big Data.

Palavras-Chave: Big Data na Cibersegurança. Violação de Dados. SIEM.

ABSTRACT

The use of Big Data in cybersecurity is an emerging solution to address the high volume of data that traditional solutions cannot handle. It can help organizations be both more effective and faster in preventing and responding to data breaches. This paper aims to explore the use of Big Data in cybersecurity, with the main goal of identifying Big Data-based cybersecurity solutions that can replace SIEM (Security Information and Event Management). To achieve the research goal, relevant concepts related to Big Data, data breaches, and SIEM were gathered to provide a better understanding of the topic. A systematic review of the relevant literature related to the topic was conducted, showcasing Big Data-based cybersecurity solutions and frameworks. Additionally, an analysis was performed on an existing case study of a company that provides Big Data-based cybersecurity services. The research goal was achieved, Big Data-based cybersecurity solutions were identified. However, it was observed that the use of Big Data does not completely replace the SIEM but rather represents an evolution of this approach. Additionally, it is relevant to highlight the existence of commercial solutions that integrate both the SIEM and Big Data analysis techniques.

Keywords: Big Data in Cybersecurity. Data Breaches. SIEM.

---

[1] Discente da Faculdade Santíssimo Sacramento, Bacharelando em Sistemas de Informação – ricardofilho11352@soumaissantissimo.com.br

[2] Docente da Faculdade Santíssimo Sacramento, especialista em Gestão da Informação (FSSS) – docente.joanmarcel@fsssacramento.br

[3] Docente da Faculdade Santíssimo Sacramento, mestre em políticas públicas e cidadania (UCSAL), especialista em gerenciamento ambiental (UCSAL), graduação em ciências sociais (UFBA) - docente.gabrielaguerreiro@fsssacramento.br

1 INTRODUÇÃO

Com o aumento de 38% no número de ataques cibernéticos globais de 2021 para 2022, evidenciado pela equipe de pesquisa da Check Point (2023), as organizações buscam se precaver contra ataques cibernéticos utilizando mecanismos de defesa cibernética para protegerem suas informações sensíveis, e evitarem as consequências negativas, apontadas por Long et al. (2017), que um vazamento de dados pode causar em uma organização, como prejuízos financeiros e possíveis danos em suas reputações, .

O fator que despertou interesse pelo tema foi a escassez de literatura científica suficiente que explorasse o uso da Big Data para prevenção de violação de dados e defesa cibernética no geral. Esse estudo visa preencher essa lacuna e destacar a importância desse tema.

Com o aumento contínuo na quantidade de dados, a rapidez com que eles são criados, e a diversidade em sua variedade - três das cinco características da Big Data apontadas por Younas (2019) - soluções tradicionais de cibersegurança, como o SIEM (em inglês, *Security Information and Event Management* ou Gerenciamento e Correlação de Eventos de Segurança, em português), são, segundo Murad et al. (2017), insuficientes para detecção de novas ameaças cibernéticas e inadequadas para tratar novas ameaças e táticas de ataque.

Dessa forma, o objetivo geral da pesquisa foi identificar soluções de cibersegurança baseadas em Big Data que possam substituir o SIEM. O SIEM foi escolhido como referência para análise, pois de acordo com Correia e Dias (2020), ele demonstra dificuldades em tratar um elevado número de dados em um período desejado de tempo.

A pergunta da pesquisa foi: Existem sistemas eficazes baseados em Big Data para prevenção de violação de dados que possam substituir a solução de segurança SIEM? Durante a pesquisa, foram encontradas soluções de segurança cibernética baseadas em Big Data, porém, constatou-se que o uso da Big Data não substitui completamente o SIEM, mas sim representa uma evolução dessa abordagem. Ademais, é importante destacar que já existem soluções comerciais que combinam tanto o SIEM quanto técnicas de análise de Big Data para fornecer serviços abrangentes de defesa cibernética. Portanto, conclui-se que o objetivo da

pesquisa foi atingido.

Este trabalho está dividido em seis capítulos, sendo o primeiro a introdução. O segundo capítulo desta pesquisa trata de cibersegurança convencional, violação de dados, e é explorado o conceito de SIEM – um exemplo de cibersegurança convencional – e os seus estágios, utilizando como arcabouço teórico o livro escrito por Miller et al. (2010) sobre SIEM. O terceiro capítulo aborda cibersegurança com Big Data, nele é explorado os principais conceitos que cercam a Big Data, como o Big Data Analytics, e mostra os estágios em comum de defesa cibernética baseada em Big Data, utilizando como referência o artigo escrito por Andrade (2020). O quarto capítulo é uma análise geral de um estudo de caso publicado pela IBM em uma empresa real com o objetivo de apresentar uma solução comercial efetiva que utiliza a abordagem baseada em Big Data. O quinto capítulo apresenta a revisão sistemática do tema, as metodologias aplicadas, os resultados, e a discussão dos resultados. Os principais autores utilizados para a revisão foram Correia e Dias (2020) e Murad, Maarof e Zainal (2017). O sexto e último capítulo aborda as considerações finais da pesquisa. Nela é mostrada a resposta para a pergunta da pesquisa e possíveis direções futuras de pesquisa em relação ao tema.

2 CIBERSEGURANÇA CONVENCIONAL

De acordo com Murad et al. (2017, p. 126), cibersegurança é um conjunto de contra-medidas, estratégias, padrões utilizados para fim de defender, detectar e prevenir quaisquer tipos de vulnerabilidades contra um sistema, rede de uma empresa ou no ciberespaço. Os autores destacam algumas das abordagens tradicionais de análise e gerenciamento de segurança, assim como os mecanismos que são geralmente utilizados em qualquer empresa ou sistemas individuais de TI, dentre elas estão a de Gerenciamento de Risco, Detecção de Malware, Detecção de Intrusão, Prevenção de Intrusão, Firewalls, Gerenciamento de Registros e o SIEM.

Este capítulo explora o conceito de violação de dados, abordagens para classificar um vazamento de dados, o conceito de SIEM, os serviços presentes na solução SIEM e seus estágios.

2.1 Violação de dados

Os dados são um dos ativos mais valiosos para uma empresa, eles podem ser convertidos em informações importantes para a tomada de decisão, além de serem úteis para o desenvolvimento de estratégias sólidas e eficazes para aplicar nas organizações (SDGGROUP, [s.d.]).

No quarto trimestre de 2020, houve aproximadamente 125 milhões de casos de violações de dados no mundo todo (SOBERS, 2022), e de acordo com o relatório *Cost of a Data Breach*, publicada pela IBM (2022), o custo total médio global de uma violação de dados é de $4,35 milhões.

Segundo Long et al. (2017, p. 2), existem duas abordagens para classificar ameaças de vazamento de dados: o vazamento de informações sensíveis de forma proposital ou não, e ameaças internas ou externas. Violações de dados realizadas por ameaças externas são causadas por invasões de hackers, *malwares,* vírus ou engenharia social.

Para Miller (2010, p. 25), ameaças externas incluem o ataque manual e humano – que pode ser desde um leigo que quer ser hacker até um grupo concentrado e organizado de hackers profissionais -, onde o atacante explora de forma sistemática e faz uso de ataques programáticos em seu sistema como vírus, *worms,* entre outros. Ataques manuais são mais lentos quanto sua progressão e mais sutis e concentrados.

Long et al. (2017, p. 1) afirmaram que a perda de informações sensíveis pode acarretar em consequências negativas para uma organização, como danos expressivos na reputação, prejuízos financeiros, e em casos mais graves pode afetar a estabilidade da empresa. Portanto, torna-se necessário buscar e/ou desenvolver métodos para aplicá-los no plano organizacional e, consequentemente, evitar danos desastrosos ou irreparáveis.

2.2 SIEM (Gerenciamento e Correlação de Eventos de Segurança)

Miller et al. (2010) apresentam o sistema SIEM como uma coleção complexa de tecnologias projetadas para fornecer uma visão holística do sistema de TI corporativo, o que beneficia tanto analistas de segurança quanto administradores de TI.

A função do SIEM, de acordo com Dias e Correia (2020, p. 293), é coletar e gerenciar dados relevantes para segurança de diferentes dispositivos em uma rede, como exemplo: *firewalls*, servidores de autenticação e entre outros. Além de

coletar e gerenciar os dados, ele vai fornecer uma maior visibilidade de segurança de rede agregando e filtrando alarmes, ao mesmo tempo que provê informação acionável para analistas de segurança.

Um dos principais objetivos do analista de segurança que faz uso do SIEM é reduzir o número de alertas falso-positivo, que é algo que acontece com certa frequência em sistemas de segurança menos sofisticados como o IDS (Sistema de Detecção de Intrusão), um número alto de alertas falso-positivo pode desperdiçar o tempo e energia do analista de segurança e fazer com que ele desconsidere os alertas positivo. Com o sistema SIEM, que é mais sofisticado, é feita a criação cuidadosa de filtros e regras de eventos correlatos para identificar e alertar apenas sobre os eventos de segurança altamente qualificados enquanto desconsidera de forma precisa o volume de eventos falso-positivo, tudo isso para diminuir ocorrências de alertas falso-positivo (MILLER et al., 2010).

Segundo Miller et al. (2010), o sistema SIEM fornece a seguinte coleção de serviços: Gerenciamento de *log* (em português, registros), Conformidade regulamentar de TI (*Compliance*), Correlação de eventos, Resposta ativa e Segurança do endpoint.

2.3 Estágios de um sistema SIEM

Miller et al. (2010, p. 92) afirmam que o SIEM é composto de várias partes e cada uma dessas partes faz um trabalho separado. Para um analista de segurança gerenciar bem um sistema SIEM e resolver problemas na medida que eles surgem, é necessário compreender cada parte do SIEM, o que cada peça faz e como funciona.

A primeira “parte” de um sistema SIEM é o dispositivo de origem que alimenta a informação para o SIEM. O dispositivo de origem é o dispositivo, aplicação, ou algum outro tipo de dado de onde você quer recuperar registros para processar e armazenar no SIEM. Sendo assim, o dispositivo de origem pode ser dispositivos como um roteador, switch, ou algum tipo de servidor, pode ser registros de uma aplicação, ou qualquer tipo de dado que é adquirido. O dispositivo de origem não é exatamente uma parte do sistema SIEM, mas é uma peça vital para o processo SIEM como um todo. (MILLER, 2010, p. 78).

A próxima etapa na anatomia do SIEM é transferir os registros do dispositivo de origem para o SIEM. Escolhe-se um dos dois métodos

fundamentais de coleção, o método de empurrar (*push method*) ou o método de puxar (*pull method*). No método de empurrar o dispositivo de origem envia seus registros para o SIEM, enquanto no método de puxar o SIEM recupera os registros do dispositivo de origem (MILLER, 2010, p. 81).

Após os registros serem enviados para o SIEM, eles estão em seu formato nativo e não tem uma boa legibilidade. Esses registros precisam ser reformatados para um formato padrão único que seja útil para o SIEM, esse processo chama-se normalização (MILLER, 2010, p. 84).

O mecanismo de regras (*rule engine*) expande a normalização de eventos de diferentes origens para disparar alertas dentro do sistema SIEM por conta de condições específicas nos registros. No início, a forma como as regras do SIEM são escritas é simples, porém, depois se torna mais complexo. Usualmente, se escreve as regras utilizando lógica booleana para determinar se condições específicas são correspondidas e examinar a correspondência de padrões dentro dos campos de dados (MILLER, 2010, p. 86-88).

O mecanismo de correlação é um subconjunto do mecanismo de regras, ele combina múltiplos eventos padrões de diferentes origens, transformando-os em um único evento correlacionado. Para o SIEM, o mecanismo de correlação agrupa eventos individuais, que podem fazer parte de um potencial incidente malicioso, em um único evento exibido no console de um operador monitorando o ambiente (MILLER, 2010, p. 86-88).

A forma de trabalhar com os volumes de registros que entram no SIEM é através do armazenamento deles para fim de retenção e consultas históricas. O modo mais comum de fazer o armazenamento é por meio dos bancos de dados, pois, eles permitem a fácil interação e recuperação de dados armazenados e por conta do seu bom desempenho ao acessar registros no banco de dados (MILLER, 2010, p. 90).

O último estágio na anatomia de um SIEM é o de monitoramento. O SIEM terá um console de interface que será baseado em navegador ou desktop, esse console será utilizado para gerenciar o SIEM. Ambas interfaces permitirão a interação com os dados armazenados no SIEM. A visualização da informação que o SIEM reuniu, pelo pessoal encarregado de tratar os incidentes, é muito mais fácil, pois, os registros foram normalizados e estão legíveis. Dentro do gerenciamento do SIEM e do console de monitoramento, o analista de segurança

poderá desenvolver o conteúdo e as regras que serão utilizadas para recuperar a informação de eventos sendo processados, o console será sua principal forma de interagir com os dados que estão armazenados no SIEM (MILLER, 2010, p. 91).

3 CIBERSEGURANÇA COM BIG DATA

De acordo com a AV-TEST, existem mais de um bilhão de *malwares* no mundo atualmente, e até o início do mês de março de 2023 já foram detectados mais de 13 milhões de novos *malwares*. Murad et al. (2017, p. 124) afirmam que as técnicas analíticas de cibersegurança existentes como análise de eventos de *log*, sistemas de detecção de intrusão e entre outros, são inadequados e não tem um bom funcionamento em grandes escalas, e tipicamente acionam alarmes falso-positivo ou falso-negativo.

Soluções tradicionais de cibersegurança como SIEM adotadas nas décadas passadas demonstram limitações ao processar Big Data, e apresentam ainda mais dificuldade em extrair as informações que esses dados podem fornecer, e isso faz com que os pesquisadores dêem mais atenção para novas técnicas para tratar um alto volume de dados relevantes para a segurança, junto com abordagens utilizando *Machine Learning* (DIAS, Luis; CORREIA, Miguel, 2020, p. 292-293).

Este capítulo aborda os conceitos principais de Big Data, o que é Big Data Analytics e como ele está relacionado com a cibersegurança, e as etapas gerais observadas em sistemas de defesa cibernética com Big Data.

3.1 Conceitos Principais de Big Data

Big Data são conjuntos de dados que são muito largos ou complexos para serem tratados por softwares tradicionais de processamento de dados. Younas (2019, p. 105) afirma que Big Data tem cinco características ou 5vs, são eles: Volume, Velocidade, Variedade, Veracidade e Valor.

Volume se refere a quantidade enorme de dados que vêm sendo gerados, coletados e processados nos últimos anos. Velocidade é a rapidez com que os dados são gerados, processados e enviados entre diferentes sistemas e dispositivos. Variedade está inserida nas características da Big data por conta dos diferentes tipos de dados e metadados que podem ser usados para os mais variados fins. Veracidade é a qualidade dos dados, tais como consistência,

confiança, segurança, confiabilidade e precisão. Valor são os diferentes tipos de benefícios que podem derivar do processamento e análise da Big Data, como por exemplo valor monetário, valor social, valor acadêmico, valor de pesquisa, entre outros (YOUNAS, 2019, p. 105).

Nyarko (Tabassum e Tyagi, 2016; Yosepu et al, 2015; Moorthy et al, 2015; Toshniwal et al, 2015; Khan et al, 2014; Shirudkar e Motwani, 2015; Ularu et al, 2012; Moura e Serrão, 2015; Hima Bindu et al, 2016 apud NYARKO, 2018, p. 17) ressalta que existem três tipos de dados compostos da *Big data*, são eles: Dados Estruturados, que são altamente organizados e em grande parte gerenciados pelo SQL, alguns exemplos são tabelas de bancos de dados, relatórios, tabelas, etc. Dados Semi Estruturados, que são gerenciados pelo XML, JSON, entre outros, é um tipo de estrutura de dados que não se adequa a uma estrutura formal, além de não possuir uma estrutura de modelo de dados. E Dados Não Estruturados, que são geradas por máquinas ou seres humanos, como mensagens de textos, *emails*, filmes, áudios, fotos, transações financeiras, *tweets*, entre outros.

De acordo com Ruban (2020), alguns métodos de coleta de dados incluem a análise de *marketing* online, que envolve a interação entre empresa e cliente, onde o cliente preenche formulários acerca de informações pessoais, as empresas utilizam desses dados para criar estratégias para melhorar o serviço ao consumidor e aumentar nas vendas. Um outro método é a coleta de dados das atividades dos usuários nas redes sociais, que são um dos principais fornecedores de dados não estruturados, onde os usuários compartilham esses dados de forma consentida.

Pratt (2022) e Ruban (2020) citam algumas formas de como os dados da Big Data podem ser coletados, dentre eles: através de *cookies*, pesquisas com os consumidores, rastreadores de *email*, empresas que vendem serviços de API, compra de dados de empresas especializadas em vender dados como a CoreLogic e Equifax, entre outros.

3.2 Big Data Analytics

Big Data Analytics se refere ao processo de coletar, examinar e analisar grandes quantidades de dados para trazer informações acerca de tendências no mercado, percepções e padrões que possam auxiliar na tomada de decisão, na maioria das vezes, de negócios. Contudo, este sub-tópico explora o uso de Big

Data Analytics na cibersegurança. Neste artigo, será utilizado, algumas vezes, o termo “técnicas/ferramentas/tecnologias de análise de Big Data” para se referir a Big Data Analytics.

Lidong e Jones (2017, p. 29) estabelecem que o uso de técnicas de análise de Big Data para mitigar problemas de segurança de rede e detecção de intrusão têm ganhado mais ênfase, pois, ele promove o estudo de fontes e formatos variados de dados para detecção de anomalias e combate à ataques cibernéticos. Eles também destacam que tecnologias da Big Data como o ecossistema Hadoop e processamento de fluxo podem armazenar grandes conjuntos de dados em alta velocidade e ajudar na realização de análise de segurança de forma mais eficiente e eficaz.

Kabanda (2021, p. 3) afirma que Big Data Analytics tem potencial para oferecer uma variedade de dimensões de segurança, como gerenciamento de tráfego de rede, padrões de acesso em transações na Internet, configuração de servidores de rede, fontes de dados de rede e credenciais de usuário. Todavia, ele ressalta que existe uma falta de conhecimento profundo e técnico ao se tratar de conceitos de Big Data Analytics como Hadoop, análise preditiva, análise de cluster e entre outros. Ele também afirma que há uma falta de infraestrutura que suporte essas inovações, e falta de cientistas de dados e políticas ou leis que promovam essas inovações.

Murad et al. (2017) inferem que sistemas de detecção de fraude podem ser considerados exemplos de uso de tecnologias de análise de Big Data na cibersegurança, por conta da alta quantidade que esses sistemas geram. Eles também mencionam que outras áreas de aplicação de detecção de fraude são as de empresas de cartão de crédito, saúde, telecomunicações e entre outros.

3.3 Estágios da defesa cibernética com Big Data

Utilizar Big Data para fins de segurança cibernética, mais especificamente prevenção de violação de dados, é um conceito relativamente novo. De acordo com Andrade (2020, p. 21), diversas arquiteturas de sistema de cibersegurança baseadas em Big Data estão sendo criadas e testadas.

Na generalidade, grande parte desses sistemas de defesa baseados em Big Data têm etapas em comum, sendo elas a etapa de coleta de dados, a de processamento de dados e a de análise de dados (Ahn e colab., 2014, apud

ANDRADE, 2020, p. 21).

De forma resumida e breve, na etapa de coleta de dados, dados de variadas fontes e distintas características são usados no “processo de ingestão de dados em um sistema de análise baseado em *Big Data*” (Ji e colab., 2016, apud ANDRADE, 2020, p. 22).

Na etapa de processamento de dados é realizada a filtragem inicial dos dados, isso acontece através de um processo que certifica se os dados atendem requisitos previamente estipulados (ANDRADE, 2020, p. 22). Por conta do grande volume de dados que são tratados, essa etapa é mais rápida e eficiente com a utilização de sistemas distribuídos ou computação em nuvem (Ahn e colab., 2014, ANDRADE, 2020, p. 23).

Na etapa de análise de dados, utiliza-se os dados estruturados da etapa anterior e, aliados a algoritmos de predição, classificação e associação, verifica se uma determinada atividade de rede é potencialmente danosa (ANDRADE, 2020, p. 28).

4 CASO DE EMPRESA REAL QUE UTILIZA BIG DATA NA CIBERSEGURANÇA

Em um estudo de caso publicado pela IBM (2022), uma empresa do Vietnã chamada Novaland que atua no ramo de corretor de imóveis, começou a notar o aumento nas ameaças cibernéticas e as vulnerabilidades nos dispositivos finais (*endpoint devices*, como computadores, laptops, etc.) dos seus funcionários, e que ferramentas de segurança geram alertas falso positivo com frequência, decidiu implementar uma solução SIEM, que é fundamental para o gerenciamento de cibersegurança. Após avaliar as melhores opções no mercado, eles escolheram o QRadar SIEM da IBM, pela automação de análises de informação de segurança, por detectar ameaças rapidamente, e pela privacidade, segurança de dados e resiliência de negócios que ele oferece.

Para implantar o QRadar SIEM na Novaland, a assistência de Serviços de Segurança da IBM ajudou a equipe de segurança cibernética da Novaland. A equipe então utilizou a ferramenta para melhorar os procedimentos e cenários de resposta a incidentes, otimizou as regras para identificar sinais de ataque e desenvolveu um conjunto de procedimentos para respostas a incidentes de segurança. O QRadar SIEM utiliza técnicas de análise inteligentes, o que contribui para a redução de alertas falso-positivo, como exemplo, a Novaland reduziu o

número de aproximadamente mil alertas falsos-positivos por dia para menos de cem, diminuindo a carga de trabalho da equipe de cibersegurança e fazendo-os focar em ameaças mais perigosas, e consequentemente tendo um aumento na produtividade.

Quadro 1: Visão geral do estudo de caso

Empresa	Novaland
Ramo	Imobiliária e investimento
Produto implantado	QRadar SIEM
Problema	Desafios crescentes de segurança cibernética devido à transformação digital e ao aumento do número de pontos vulneráveis durante a pandemia da COVID-19
Desafios	Integração de dados de segurança e geração de falsos positivos pelos sistemas de segurança, dificultando a priorização e resolução eficiente das ameaças
Solução	Implementação do QRadar SIEM e otimização dos procedimentos de resposta a incidentes e cenários
Benefícios	Detecção de ameaças acelerada, resposta rápida a incidentes, redução de falsos positivos, priorização eficiente de ameaças, proteção aprimorada de sistemas, informações de clientes e propriedade intelectual, eficiência operacional e melhoria na confiança de investidores e clientes

Fonte: O autor

5 REVISÃO SISTEMÁTICA

Este capítulo está dividido em metodologia, onde são apresentadas técnicas utilizadas para realização da revisão sistemática, resultados, onde é revelado os achados mais relevantes e importantes dos artigos selecionados levando em consideração as hipóteses formuladas e o objetivo geral da pesquisa, e por último a discussão dos resultados.

5.1 Metodologia

Em relação a tipologia de pesquisa, o objetivo tem caráter exploratório, buscando tornar o tema o mais familiar e claro possível, quanto aos procedimentos técnicos ela foi bibliográfica, fazendo uso de artigos científicos e trabalhos experimentais relacionados ao tema e os conceitos estudados, e foi feita uma revisão sistemática de artigos relacionados ao uso de Big Data na Cibersegurança.

O método de abordagem foi o hipotético-dedutivo, ou seja, as hipóteses foram testadas, e uma dedução foi feita partindo de termos gerais para mais específicos com o objetivo de verificar se as hipóteses são verdadeiras ou falsas.

O método de procedimento foi o monográfico, visto que o tema exige estudo e investigação exaustiva e detalhada para conseguir atingir seu objetivo geral. Foram utilizados na pesquisa dados do tipo secundário, neste caso, materiais que receberam tratamento analítico, como tese, artigos e trabalho experimental.

A revisão sistemática foi feita através da busca de artigos no Google Acadêmico e ResearchGate, utilizando os seguintes descritores: “Big Data na Cibersegurança”, “Big Data in Cybersecurity”, “Big Data Analytics in Cybersecurity”, “Big Data and SIEM”. Os critérios de inclusão foram: artigos escritos no idioma português ou inglês publicados a partir de 2017, para incluir as pesquisas e desenvolvimentos recentes no tema; artigos que relacionam a solução SIEM no contexto da Big Data; e artigos que abordam ou apresentam soluções ou arquiteturas de cibersegurança baseadas em Big Data. Os critérios de exclusão foram: Artigos que não estão disponíveis em formato acessível, como artigos pagos ou sem acesso público; Artigos que discutem apenas os benefícios e potenciais do uso da Big Data na Cibersegurança. Para a seleção de um artigo, ele deve ter atendido pelo menos um critério de inclusão, contanto que não tenha violado nenhum critério de exclusão.

Quadro 2: Artigos selecionados para a pesquisa

TÍTULO	AUTOR / ANO	TIPO DE ESTUDO
Big Data Analytics Adoption for Cybersecurity: A Review of Current Solutions, Requirements, Challenges and Trends	Murad, Maarof e Zainal, 2017	Artigo de revisão
Big Data Analytics for Intrusion Detection: An Overview	Correia e Dias, 2020	Artigo de revisão
Big data in cybersecurity: a survey of applications and future trends	Alani, 2021	Artigo de revisão
O uso da Big Data na prevenção de ataques cibernéticos	Andrade, 2020	Pesquisa Exploratória
Smart SIEM: From Big Data logs and events to Smart Data alerts	Arass e Souissi, 2019	Pesquisa Experimental

Fonte: O autor

Quadro 3: Verificação dos critérios de inclusão dos artigos selecionados

Autor / Ano	**No idioma português ou inglês publicados a partir de 2017**	**Relacionam a solução SIEM no contexto da Big Data**	**Apresentam soluções ou arquiteturas de cibersegurança baseadas em Big Data**
Murad, Maarof e Zainal, 2017	Atende critério	Atende critério	Atende critério

Correia e Dias, 2020	Atende critério	Atende critério	Atende critério
Alani, 2021	Atende critério	Não atende critério	Atende critério
Andrade, 2020	Atende critério	Não atende critério	Atende critério
Arass e Souissi, 2019	Atende critério	Atende critério	Atende critério

Fonte: O autor

5.2 Resultados

Neste tópico são apresentados os resultados obtidos a partir dos dados coletados nos artigos selecionados. O primeiro subtópico relaciona SIEM e Big Data no contexto da Cibersegurança, traz desafios na aplicação de Big Data na Cibersegurança encontrados nos artigos, e resume os artigos selecionados que abordam essas duas variáveis. O segundo subtópico levanta algumas soluções e arquiteturas de cibersegurança baseadas em Big Data presentes nos artigos selecionados.

5.2.1 SIEM e Big Data na Cibersegurança

No artigo de Murad, Maarof e Zainal (2017), os autores apontam as deficiências do SIEM e propõem que uma solução de cibersegurança baseada em Big Data pode superar essas deficiências. Os autores apresentam os estágios de evolução dos sistemas de detecção de intrusão denominados pela CSA (*Cloud Security Alliance*), a primeira geração ou estágio é o Sistema de Detecção de Intrusão, a segunda geração é o SIEM, e a terceira geração é o Big Data Analytics (a 2º geração do SIEM). Na terceira geração, ocorre o avanço dos esforços feitos na 2º geração com o uso de Big Data Analytics, reduzindo o consumo de tempo na correlação e consolidação de informação de eventos de segurança. Eles levantam os requisitos (tratar dados de múltiplas fontes, gerenciamento de dados em grande escala, visualização de dados e tecnologia de infraestrutura de alto desempenho) para adotar tecnologias de análise de Big Data em cibersegurança, para superar as fraquezas de sistemas de cibersegurança tradicionais e baseados em SIEM. Sistemas de cibersegurança baseados em big data são levantados neste trabalho.

Murad, Maarof e Zainal (2017) também destacam as dificuldades e preocupações em adotar Big Data Analytics para cibersegurança, dentre elas, como lidar com dados não estruturados, a questão da privacidade dos dados

utilizados, a origem dos dados utilizados, e como aplicar de forma otimizada análise de dados em tempo real para defesa cibernética. Os autores concluem que ferramentas tradicionais de cibersegurança - inclusive o SIEM - são inadequadas para tratar novas ameaças e táticas de ataque, e que essas ferramentas não são suficientes para a detecção de ciber-ameaças, por conta do alto volume e vários tipos de dados que são gerados atualmente. Segundo os autores, é necessário implementar ferramentas de análise de Big Data para conter ameaças sofisticadas.

A pesquisa de Correia e Dias (2020) é uma revisão do estado-da-arte acerca de novas técnicas para tratar um alto volume de dados relevantes para segurança, junto com a abordagem de Machine Learning. Os autores apresentam algumas soluções de cibersegurança baseada em big data com machine learning que se complementam com o SIEM que estão presentes na literatura científica.

Correia e Dias (2020), assim como no artigo de Murad, Maarof e Zainal (2017), também mostram os estágios de evolução dos sistemas de detecção de intrusão, e se referem a 3º geração (Big Data Analytics) como a 2º geração do SIEM. Segundo os autores, técnicas para tratar Big Data são necessárias para o domínio de cibersegurança, por conta do grande volume de dados que as plataformas SIEM não conseguem tratar, pelo menos não em um período desejado de tempo. Dentre os desafios citados para a implementação de Machine Learning e uso da Big Data na cibersegurança, são destacados a evolução dos ataques cibernéticos, o desempenho de detecção de intrusão, as capacidades para análise de dados em tempo real e as preocupações com a privacidade e integridade dos dados que são utilizados.

Correia e Dias (2020) afirmam que os principais fornecedores de SIEM estão evoluindo para o uso de tecnologias de computação distribuída e estão integrando Machine Learning em suas soluções, porém, reconhecem que o uso de Machine Learning na cibersegurança ainda precisa ser mais explorado, e que ao aplicar essa técnica no domínio de cibersegurança, o maior desafio é o fator humano que chega a ser imprevisível, o que torna difícil para um sistema automatizado de fazer predições precisas. Eles concluem que uma solução automatizada para detecção de intrusão pode ser díficil de alcançar, mas afirmam que Machine Learning é o melhor complemento para alcançar uma conscientização contextual de ameaças próxima ao tempo real.

Alani (2021), em seu artigo, reúne as pesquisas mais atuais em diferentes campos de aplicações de Big Data na Cibersegurança, dentre elas, detecção de anomalia e intrusão, detecção de *ransomware* e *malware*, e segurança na nuvem. Alani também aponta possíveis direções futuras para pesquisa de aplicações da Big Data na cibersegurança. Para o autor, aproveitar os benefícios da Big Data ao construir sistemas de cibersegurança robustos, adaptáveis e rápidos, é mais uma necessidade do que uma escolha. Ele afirma que os métodos convencionais de detecção são inferiores em termos de precisão e capacidade para detectar ameaças, por conta do volume substancial de dados.

Apesar de Alani (2021) reforçar que sistemas de detecção de ameaças que não fazem uso de tecnologias de análise de Big Data, Machine Learning, e computação na nuvem, podem rápido e facilmente ficarem obsoletos, e que o futuro da cibersegurança está conectado com Big Data, ele alerta em seu artigo para ter precaução e atenção ao utilizar machine learning e Big Data em aplicações de cibersegurança. Ele finaliza, mostrando as possíveis direções de pesquisa acerca do uso de Big Data em aplicações de Cibersegurança, dentre elas está Detecção de Intrusão, Detecção de Fraude e Detecção de Malware.

5.2.2 Soluções e Arquiteturas de Cibersegurança baseadas em Big Data

Arass e Souissi (2019), em seu artigo, propôs um SIEM de código-aberto de nova geração composto pela plataforma de Big Data ELK – para monitorar atividade maliciosa - e integrada com outras ferramentas de detecção de intrusão e balanceamento de carga - para otimizar o protótipo em tempo real -, os autores denominaram o sistema de Smart SIEM. O Smart SIEM deve ser adaptado para o contexto da Big Data para monitoramento da segurança de dispositivos de TI. Os autores identificaram que a maioria dos SIEMs de alto desempenho são demasiado caros e existe pouco conhecimento sobre sua arquitetura interna, por não serem divulgados para os usuários, e por esses SIEMs de alto desempenho serem caros, poucas organizações adquirem eles. Eles também notaram que os SIEMs clássicos não tratam Big Data.

Arass e Souissi (2019) destacam que a plataforma ELK não é uma solução SIEM por si só, todavia eles combinaram esta plataforma com outras ferramentas de segurança para atender aos requisitos de sistema da nova geração do SIEM. O ELK é composto pelas seguintes tecnologias: Beats (envia registros de dados

para o Logstash), Logstash (coleta e integra dados de diversas fontes ao mesmo tempo e envia para um sistema de armazenamento como o Elasticsearch), Elasticsearch (permite a pesquisa e análise de dados em tempo real, e utiliza um banco de dados não relacional) e Kibana (uma ferramenta potente e intuitiva de visualização de dados que são encontrados no Elasticsearch).

As ferramentas utilizadas por Arass e Souissi (2019) para integrar com o ELK foram o Snort (um sistema que detecta e previne intrusões na rede), Sguil (fornece interface gráfica intuitiva para visualizar eventos, dados de sessão e capturas de pacotes brutos), Zeek (fornece uma plataforma para análise de rede), OSSEK (um sistema de detecção de intrusão que gera alertas em tempo real) e Redis (para fornecer monitoramento em tempo real, balanceamento de carga e gerenciamento de fila). Para a validação do Smart SIEM, foram executadas uma série de ações cibernéticas maliciosas utilizando a máquina Linux Kali, em um laboratório virtual, para avaliar o quão bem o protótipo estava detectando e relatando essas ações. De acordo com os autores, o Smart SIEM respondeu perfeitamente aos ataques feitos pelo Linux Kali, conseguindo detectar ataques de injeção de comando SQL, ataques de força bruta e ataques de DDoS, e através do Smart SIEM eles conseguiram identificar o alvo e as máquinas do atacante. Eles concluem que graças ao Smart SIEM, o analista pode, além de monitorar a atividade de cibersegurança de sua rede, agilizar o processo de detecção e respostas a incidentes de segurança.

No artigo de Murad, Maarof e Zainal (2017), dentre as soluções de cibersegurança que eles mencionaram estão inclusas o QRadar e Infosphere BigInsights da IBM e a Plataforma de Inteligência de Segurança da LogRhythm. As soluções QRadar são utilizadas por grandes empresas para reunir e correlacionar bilhões de eventos e tráfego de redes todos os dias. O QRadar também inclui o SIEM, sistemas de detecção de anomalias, sistemas de gerenciamento de registros, sistemas forenses e sistemas de gerenciamento de vulnerabilidades e configuração em uma arquitetura unificada.

Murad, Maarof e Zainal (2017) afirmam que a IBM desenvolveu o InfoSphere BigInsights como uma extensão do QRadar, enquanto o QRadar utiliza dados derivados de fontes tradicionais como dados de usuário, de rede, e de atividades da aplicação, o InfoSphere utiliza dados não estruturados baseados em infraestruturas de plataforma para Big Data como o Hadoop, para executar

técnicas de análise customizadas e melhorar a detecção de ameaças em tempo real. A integração destas duas tecnologias resulta em “uma solução inteligente que reúne, monitora, analisa, detecta e relata qualquer incidente de cibersegurança que possa ocorrer em tempo real” (tradução nossa).

A solução da LogRhythm é, de acordo com Murad, Maarof e Zainal (2017), um SIEM da 2º geração, pois utiliza técnicas de análise de Big Data na cibersegurança. “Ele combina os processos de gerenciamento de registros, monitoramento de rede, forense digital e gerenciamento de ameaças em uma única plataforma, além de fornecer procedimentos para resposta em tempo real” (tradução nossa). Os autores destacam também que, segundo os projetistas desta plataforma, esta solução também fornece um mecanismo de detecção precoce do comportamento de ataques, fazendo com que previna violações antes de elas acontecerem.

No artigo de Alani (2021), ele revisou diversas propostas de cibersegurança baseada em Big Data de outros autores em diversas aplicações. Na aplicação na área de detecção de intrusão e anomalias, uma das propostas mencionadas foi a do autor Kotenko et al. (2020), onde ele propõe uma abordagem utilizando machine learning e tecnologias da Big Data. A abordagem proposta emprega duas diferentes arquiteturas de IDS distribuído baseadas em Big Data, os testes nesta abordagem foram feitos utilizando dois grandes conjuntos de dados, o primeiro conjunto foi tráfego de Internet das Coisas incluindo diversos tipos de ataques e o segundo conjunto foi tráfego de redes de computadores incluindo ataques DDoS. Os resultados foram satisfatórios em termos de precisão e rapidez de detecção.

Andrade (2020) cita quatro propostas e explica o funcionamento de cada uma das arquiteturas. Uma das arquiteturas mencionadas é a de Razaq e colab. (2016) e tem como tecnologias o *MySQL*, *Hadoop Sqoop* e *HDFS*. O sistema coleta dados de fontes distintas, faz o armazenamento temporário deles no *MySQL*, então os dados são transferidos de forma definitiva para o *HDFS*, depois esses dados armazenados no *HDFS* são processados, “criando vetores de características úteis para identificar anomalias”, e finalmente os dados são analisados para identificar observações anômalas que podem ser ciberataques genuínos em potencial.

5.3 Discussão

Existem três estágios evolutivos no domínio de detecção de intrusão, o Big Data Analytics é a terceira geração deste estágio, e ele também é considerado como a 2º Geração do SIEM (Murad, Maarof e Zainal, 2017; Correia e Dias, 2020). Diante disso, o uso de técnicas de análise de Big Data é uma evolução do SIEM.

Artigos que mencionam softwares utilizados em soluções baseadas em Big Data, como o Smart SIEM no artigo de Arass e Souissi (2019) e propostas apresentadas no artigo de Andrade (2020), utilizam ferramentas de análise de Big Data. Isso indica que a utilização do Big Data Analytics é um requisito fundamental para constatar que uma solução ou arquitetura faz uso de Big Data para cibersegurança.

Murad, Maarof e Zainal (2017) afirmaram que utilizar apenas o SIEM é insuficiente para tratar o grande volume de dados que vêm sendo criados constantemente e detectar ameaças cibernéticas. Nos artigos selecionados foram notadas soluções de defesa cibernética recentes que utilizam Big Data e SIEM e soluções que utilizam Big Data, mas sem menção do SIEM.

Percebe-se, pelos artigos e estudo de caso apresendado no capítulo 4, que existem atualmente soluções comercializadas de empresas de cibersegurança que utilizam Big Data na defesa cibernética, como o QRadar SIEM da IBM e o LogRhythm SIEM da LogRhythm.

No artigo de Andrade (2020), ele afirma que é possível construir um sistema de defesa cibernética sem precisar pagar licenciamento de software especializado, as propostas que ele citou foram a de Campiolo e colab. (2018), Razaq e colab. (2016), Shenwen e colab. (2015) e Klein e colab. (2016). Todavia, para validar se essas propostas poderiam ser utilizadas em uma infraestrutura de cibersegurança ou para comercialização, seria necessário uma série de testes, para avaliar o monitoramento em tempo real, o quão rápido é a detecção de intrusão, entre outros aspectos.

O Smart SIEM, proposto por Arass e Souissi (2019), assim como nas propostas do artigo de Andrade (2020), é composto por ferramentas que não possuem custo de licenciamento. Testes foram feitos para validar esta solução e obteve resultados satisfatórios. Perante o exposto, é possível construir uma solução de cibersegurança baseada em Big Data sem ter de arcar com custo de

licenciamento, contudo, ainda é inconclusivo afirmar que estas soluções estejam prontas para serem utilizadas para fins comerciais, para validar isso, mais estudos e mais testes devem ser feitos.

No projeto de pesquisa, foram feitas três hipóteses para testar, se elas são verdadeiras ou falsas, ao obter os resultados da pesquisa. A primeira hipótese: Soluções baseadas em Big Data de prevenção a ataques cibernéticos são mais eficientes do que a solução de segurança SIEM; A segunda hipótese: Implementações de soluções de prevenção de violação de dados baseadas em Big Data são menos custosas que a implementação da solução SIEM; A terceira hipótese: A solução SIEM está ficando obsoleta, portanto, em breve será substituída por soluções de cibersegurança baseadas em Big Data.

A primeira hipótese se provou falsa, pois, soluções de prevenção a ataques cibernéticos baseadas em Big Data apresentam eficiência equiparada à solução SIEM, sendo ambos complementares em uma abordagem de defesa cibernética, como é o exemplo do Smart SIEM e a Plataforma de Inteligência de Segurança da LogRhythm mostrados nos resultados.

A segunda hipótese é inconclusiva. Foram apresentados nos resultados, duas soluções de defesa cibernética que utilizam Big Data - o Smart SIEM e a proposta de Razaq e colab. (2016) mencionado por Andrade (2020) -, ambas utilizando licenciamento de software gratuito em suas arquiteturas, o que leva-se a entender que estas soluções são pouco custosas. Todavia, não foram encontrados nos artigos selecionados os custos de uma implementação de uma solução SIEM tradicional - apesar de ser mencionado que é caro - para fazer um comparativo. Além disso, seria necessário dados específicos acerca dos custos não só de software, mas de hardware e infraestrutura de soluções baseadas em Big Data para fazer uma comparação completa e precisa.

A terceira hipótese é parcialmente verdadeira, pois, o uso exclusivo do SIEM tradicional é insuficiente para tratar a grande quantidade diversificada de dados que são gerados atualmente. Contudo, esta solução não será substituída por soluções de cibersegurança baseadas em Big Data, ela está sendo evoluída para se adaptar ao cenário da Big Data, ou seja, a solução SIEM atualmente utiliza técnicas de análise de Big Data, como exemplo o QRadar, conforme apresentado nos resultados, para proporcionar benefícios como detecção avançada de ameaças e redução de alertas falso-positivo. Pode-se concluir que

Big Data Analytics e SIEM são complementares e podem ser integrados em uma solução de cibersegurança.

6 CONSIDERAÇÕES FINAIS

Com o aumento crescente de casos de violação de dados e com a percepção de que ferramentas tradicionais de cibersegurança, como o SIEM, não conseguem tratar o alto volume de dados de diversos tipos que são gerados atualmente, tornou-se necessário que as empresas que oferecem serviços de cibersegurança passassem a utilizar tecnologias de análise de Big Data (Big Data Analytics) em suas soluções de defesa cibernética, para que seja possível tratar estes dados de maneira rápida e eficiente.

As ferramentas de análise de Big Data possuem grande potencial para serem utilizados para diversas aplicações de defesa cibernética, e atualmente, existem empresas que utilizam destas ferramentas para cibersergurança, como a IBM e a LogRhythm.

Existem três estágios evolutivos no domínio de detecção de intrusão, o Big Data Analytics é a terceira geração deste estágio, e é considerado a segunda geração do SIEM. Com isso, torna-se evidente que o uso de ferramentas de análise de Big Data representa uma evolução da solução SIEM.

Com base no objetivo geral da pesquisa, podemos concluir que ele foi alcançado, visto que foi apresentado nos resultados, soluções de defesa cibernética baseadas em Big Data. Todavia, foi descoberto durante a pesquisa que o uso da Big Data na cibersegurança não é uma substituição do SIEM, mas sim uma evolução desta abordagem, e que já existem soluções comerciais que integram tanto o SIEM quanto técnicas de análise de Big Data em seus serviços de segurança cibernética.

Futuras pesquisas podem explorar os desafios e precauções na implementação de uma arquitetura de cibersegurança baseada em Big Data e seus requisitos técnicos, estudos que analisem os custos envolvidos na implementação de soluções SIEM tradicionais e soluções de segurança cibernética que fazem uso da Big Data, e o desenvolvimento de métodos para proteção e privacidade dos dados sensíveis utilizados no processo de análise e processamento de Big Data. Esses tópicos podem avançar o conhecimento e aprimorar as práticas de segurança cibernética com Big Data.

7 REFERÊNCIAS

ALANI, Mohammed M. **Big data in cybersecurity**: a survey of applications and future trends. 2021.

ANDRADE, Luiz Claudio Oliveira de. **O uso do Big Data na prevenção de ataques cibernéticos**. 2020.

ARASS, Mohammed El; SOUISSI, Nissrine. **Smart SIEM**: From Big Data logs and events to Smart Data alerts. 2019.

AV-ATLAS. **ABOUT MALWARE AND PUA**. Disponível em: <https://portal.av-atlas.org/malware>. 2023. Acesso em: 5 Mar. 2023.

AV-TEST. **Malware**. Disponível em: <https://www.av-test.org/en/statistics/malware/>. 2023. Acesso em: 5 Mar. 2023.

CHECKPOINT. **Check Point Research Reports a 38% Increase in 2022 Global Cyberattacks**. Disponível em: < https://blog.checkpoint.com/2023/01/05/38-increase-in-2022-global-cyberattacks/>. 2023. Acesso em: 29 Jun. 2023.

DIAS, Luís; CORREIA, Miguel. **Big Data Analytics for Intrusion Detection:** An Overview. 2020.

IBM. **Accelerated security threat detection and priority response**. 2022. Disponível em: <https://www.ibm.com/case-studies/novaland>. Acesso em: 22 Mai. 2023.

IBM. **Cost of a data breach 2022**. Disponível em: <https://www.ibm.com/reports/data-breach>. 2022. Acesso em: 24 Out. 2022.

KABANDA, Gabriel; **Performance of Machine Learning and Big Data Analytics Paradigms in Cybersecurity and Cloud Computing Platforms**. 2021.

LIDONG, Wang; JONES, Randy; **Big Data Analytics for Network Intrusion Detection**: A Survey. 2017.

LONG, Cheng; FANG, Liu; DANFENG, Yao. **Enterprise data breach:** causes, challenges, prevention, and future directions. 2017.

MILLER, David R. et al. **Security Information and Event Management (SIEM) Implementation**. Local de publicação: McGraw-Hill Companies, 15 Nov. 2010.

MURAD, A. Rassam; MOHD., Aizaini Maarof; ZAINAL, Anazida. **Big Data**

Analytics Adoption for Cybersecurity: A Review of Current Solutions, Requirements, Challenges and Trends. 2017.

NYARKO, Richard. **Security of Big Data:** Focus on Data Leakage Prevention (DLP). 2018.

PRATT, Mary. **How big data collection works:** Process, challenges, techniques. **TechTarget**, 2022. Disponível em: <https://www.techtarget.com/searchdatamanagement/feature/Big-data-collection-processes-challenges-and-best-practices>. Acesso em: 25 Out. 2022.

RUBAN, Viktoria. **HOW IS BIG DATA COLLECTED BY COMPANIES? Computools**, 2020. Disponível em: <https://computools.com/how-is-big-data-collected/>. Acesso em: 25 Out. 2022.

SDGGROUP. **WHY IS DATA A VALUABLE ASSET?** SDGGROUP, [s.d.]. Disponível em: <https://www.sdggroup.com/en-US/insights-room/why-data-valuable-asset-0>. Acesso em: 29 Jun. 2023.

SOBERS, Rob. **89 Must-Know Data Breach Statistics [2022]**. **Varonis**, 2022**.** Disponível em: <https://www.varonis.com/blog/data-breach-statistics>. Acesso em: 24 Out. 2022.

YOUNAS, Muhammad. **Research challenges of big data**. 2019.