ESCOLA DE COMANDO E ESTADO-MAIOR DO EXÉRCITO

ESCOLA MARECHAL CASTELLO BRANCO

Maj QEM **LUIZ CLAUDIO** OLIVEIRA DE ANDRADE

O uso do Big Data na prevenção de ataques cibernéticos

Rio de Janeiro
2020

Maj QEM **LUIZ CLAUDIO** OLIVEIRA DE ANDRADE

O uso do Big Data na prevenção de ataques cibernéticos

Trabalho de Conclusão de Curso apresentado à Escola de Comando e Estado-Maior do Exército, como pré-requisito para matrícula no Curso de Especialização em Ciências Militares, com ênfase em Defesa.

Orientador: Ten Cel Com Ronaldo André Furtado

Rio de Janeiro
2020

A553u Andrade, Luiz Claudio Oliveira de

O uso do Big Data na prevenção de ataques cibernéticos. / Luiz Claudio Oliveira de Andrade. ―2020.
52 f. : il. ; 30 cm.

Orientação: Ronaldo André Furtado.
Trabalho de Conclusão de Curso (Especialização em Ciências Militares)―Escola de Comando e Estado-Maior do Exército, Rio de Janeiro, 2020.
Bibliografia: f. 47-51

1. *BIG DATA*. 2. DEFESA CIBERNÉTICA. 3. ATAQUE APT. I. Título.

CDD 003.5

Maj QEM **LUIZ CLAUDIO** OLIVEIRA DE ANDRADE

O uso do Big Data na prevenção de ataques cibernéticos

Trabalho de Conclusão de Curso apresentado na Escola de Comando e Estado-Maior do Exército, como requisito parcial para a obtenção do título de Especialista em Ciências Militares, com ênfase em Defesa Nacional.

Aprovado em _____ de_____ de_____.

COMISSÃO AVALIADORA

_____
Ronaldo André Furtado – Ten Cel Com - Presidente
Escola de Comando e Estado-Maior do Exército

_____
Luiz Adolfo Sodré de Castro Júnior – Ten Cel Cav – Membro
Escola de Comando e Estado-Maior do Exército

_____
Adriano de Paula Fontainhas Bandeira – Maj QEM – Dr. – Membro
Escola de Comando e Estado-Maior do Exército

Primeiramente, dedico esse trabalho à Deus, por ter permitido que eu chegasse até aqui.

Em segundo, à minha família, principalmente minha esposa, Waleska, por ser minha melhor amiga e companheira em todos os momentos, me dando todo o apoio necessário para finalizar mais um desafio.

RESUMO

A presente pesquisa apresenta de que maneira o *Big Data* pode ser utilizado na prevenção de ataques cibernéticos (ou ciberataques). Para tanto, foram estudadas as definições de ataque cibernético utilizadas pelo dos Estados Unidos da América (EUA) e pelos Estados membros da Organização para a Cooperação de Xangai (OCX), o que permitiu concluir que um ciberataque consiste em qualquer ação tomada com o objetivo de infligir prejuízo cibernético à parte opositora, que pode estar no nível pessoal, organizacional e até mesmo Estatal, fazendo com que sistemas e infraestruturas de rede não se comportem conforme o planejado. Nesse contexto, foi feito um estudo sobre os ataques denominados ameaças persistentes avançadas (*Advanced Persistent Threats* - APT), onde se mostrou que os mecanismos de defesa tradicionais têm dificuldade para detectar esse tipo de ataque. Da mesma forma, foi apresentado que entidades estatais têm sido priorizadas por esses ataques, figurando entre os dez principais alvos em 2019, o que torna importante a capacidade de se defender desses ataques. Posteriormente, foram caracterizadas as principais etapas da defesa cibernética baseadas em *Big Data*: coleta de dados, processamento de dados e análise de dados. Em seguida, foram apresentadas as arquiteturas de alguns sistemas, onde foram identificadas as tecnologias utilizadas. Na discussão das arquiteturas, pôde-se observar que, apesar dos sistemas possuírem as mesmas etapas, há uma grande disponibilidade de tecnologias que permite diferentes formas de arquitetar sistemas de defesa cibernética baseado em *Big Data*. Dessa forma, buscando resolver o questionamento de como escolher tecnologias e detalhes de arquitetura, foi apresentada uma arquitetura de referência para sistemas de defesa cibernética baseados em *Big Data* no domínio da segurança nacional.

Palavras-chave: *Big Data***.** Defesa cibernética**.** Ataque APT.

ABSTRACT

The present research shows how Big Data can be used to prevent cybernetic attacks (or cyberattacks). Therefore, the definitions used by the Unites States of America (USA) and by the member states of the Shanghai Cooperation Organization (SCO) were studied, which allowed to conclude that a cyberattack consists in any action taken with the objective to inflict cybernetic harm to the opposing party, which may be in the personal, organizational and even in the state level, making system and infrastructures to not behave in the planned way. In this context, a study was made on Advanced Persistent Threats (APT) attacks, where it is showed that traditional defense mechanisms have difficulty to detect such attacks. Likewise, it was presented that government entities have been prioritized by these attacks, figuring among the top ten targets in 2019, which makes the ability to defend against these attacks important. Subsequently, the main stages of cyber defense based on Big Data were characterized: data collection, data processing and data analysis. Next, the architecture of some systems was presented, where the technologies used were identified. In the discussion, it was noted that, although the systems possess the same stages, there is a great availability of technologies that allows different ways of architecting cyber defense systems based on Big Data. Thus, seeking to solve the question on how to choose technologies and architectural details, a reference architecture for cyber defense systems based on Big Data in the national domain was presented.

Keywords: *Big Data***.** Cybernetic defense**.** APT attack.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A sociedade moderna está vivendo a chamada Era da Informação. Isso significa que, a cada momento, o conhecimento se torna mais valorizado. Nesse sentido, a busca de conhecimento tem resultado em inúmeras formas de se obter informação, o que faz com que estejamos rodeados, e até mesmo imersos, em sistemas que coletam e produzem informação constantemente, gerando uma grande massa de dados que é genericamente definida como *Big Data* (RUSSOM, 2011).

O *Big Data* tem grande potencial de uso, pois essa enorme massa de dados a que temos acesso pode ser analisada na busca de padrões, na busca de correlações e na antecipação de tendências (ALVES, 2018), que podem ser utilizados na tomada de decisões. Porém, cabe ressaltar que apenas uma pequena parcela dos dados consegue ser analisada atualmente (GALLAHER, 2016), o que mostra uma fragilidade no uso do *Big Data*.

Nesse sentido, o desenvolvimento de soluções que consigam se valer do uso do *Big Data* com mais efetividade vem se tornando cada vez mais importante para organizações nos mais diversos níveis (MISHRA e SINGH, 2016). Tal fato está se tornando uma tendência mundial que está gerando influências nas políticas nacionais de defesa de diversos países.

No campo da defesa cibernética, tem se constatado que os ataques cibernéticos se tornaram uma tendência para, entre outros objetivos, coletar informações sensíveis sobre um alvo e até mesmo sabotar o funcionamento de sistemas sensíveis. Considerando que as ferramentas tradicionais de defesa cibernética têm dificuldade de identificar os ataques cibernéticos modernos (CAMPIOLO e colab., 2018), o uso do *Big Data* revela ter potencial na prevenção desses ataques (ZUECH e colab., 2015).

Esses ataques modernos são denominados ameaças persistentes avançadas (*Advanced Persistent Threats* - APT), que são ataques multifacetados, sofisticados, multifásicos e de longa duração focados em um alvo particular (ALGULIYEV e IMAMVERDIYEV, 2014). Tais ataques têm sido utilizados como armas cibernéticas e têm contribuído para a classificação da defesa cibernética como um aspecto fundamental da segurança nacional (BRASIL, 2016).

Com a premissa do uso do *Big Data,* novos sistemas de defesa cibernética podem ser criados. Tais sistemas são, de maneira genérica, focados na coleta,

processamento e análise dos dados e usam algoritmos de predição, classificação e associação para identificar ataques ou comportamentos anormais, permitindo uma ação tempestiva de defesa cibernética.

Nesse sentido, se constata, inicialmente, que o *Big Data* tem grande valor operacional na prevenção de ataques cibernéticos. Diversos estudos atuais têm comprovado a eficácia de seu uso nas mais variadas áreas da computação (LIU e colab., 2019), o que revela a importância do estudo do assunto no Exército Brasileiro (EB).

Dessa forma, considerando o potencial que o uso do *Big Data* tem na prevenção de ataques cibernéticos, se chegou ao problema deste trabalho: *De que maneira o Big Data pode ser utilizado na prevenção de ataques cibernéticos?*

A solução para o problema apresentado se materializa com o seguinte objetivo geral: *Apresentar de que maneira o Big Data pode ser utilizado na prevenção de ataques cibernéticos.*

Tal objetivo geral, para ser cumprido, é composto dos seguintes objetivos específicos:

a) estudar definições de ataque cibernético;
b) caracterizar as principais etapas da defesa cibernética baseadas em *Big Data*;
c) apresentar arquiteturas de defesa cibernética baseadas em *Big Data*; e
d) discutir as arquiteturas de defesa baseadas em *Big Data.*

2 ATAQUES CIBERNÉTICOS

Não existe uma definição única e pacificada do termo “Ataque Cibernético” ou, resumidamente, ciberataques. A falta de uma definição compartilhada do termo tem tornado difícil o desenvolvimento de políticas que embasem uma atuação governamental coordenada no combate a esses ataques (HATHAWAY e colab., 2012).

Nesse sentido, comprovando a urgência do tema da Defesa Cibernética, pode-se analisar dois esforços governamentais proeminentes para entender a ameaça imposta pelos ciberataques. O primeiro esforço é dos Estados Unidos da América (EUA) e o segundo é o esforço conjunto dos Estados membros da Organização para a Cooperação de Xangai (OCX), uma organização internacional focada na defesa conjunta, na economia e infraestrutura da China, Rússia, Cazaquistão, Quirguistão, Tajiquistão e Uzbequistão (BAILES e colab., 2007), mais recentemente Índia e Paquistão entraram na cooperação.

De acordo com a definição do Departamento de Defesa dos EUA, um ciberataque é definido como:

> Um ato hostil usando computador, redes ou sistemas relacionados e direcionado para interromper e/ou destruir funções, bens ou sistemas cibernéticos críticos de um adversário. Os efeitos pretendidos de ciberataque não são necessariamente limitados aos sistemas de computador ou dados visados—por exemplo, ataques em sistemas de computador que são focados em degradar ou destruir infraestrutura ou capacidade de C2. Um ataque cibernético pode usar veículos de entrega intermediários incluindo dispositivos periféricos, transmissores eletrônicos, código embarcado ou operadores humanos. A ativação ou efeito de um ciberataque podem ser largamente separados temporalmente e geograficamente da entrega. (JAMES E., 2011, tradução nossa)

Como se pode observar, a definição dos EUA é focada no objetivo de causar dano a sistemas cibernéticos críticos, sendo, portanto, uma definição pragmática e objetiva de ciberataques.

Por outro lado, a definição dada pela OCX é mais ampla. A organização criou uma definição que define ciberataques como sendo o uso de tecnologias da informação e comunicação (TIC) e quaisquer outros meios para afetar estruturas sensíveis e até mesmo estruturas políticas e sociais de um Estado (OCX, 2009). Sendo assim, a definição da OCX inclui uma visão mais ampla, abrangendo o uso

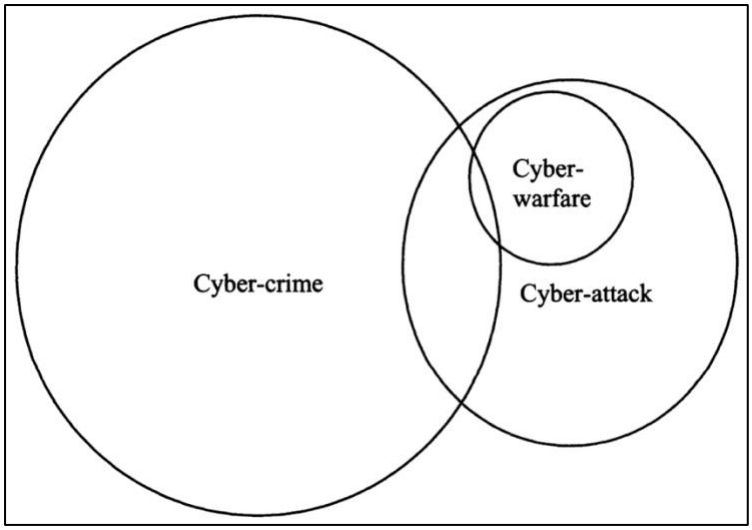
de tecnologia cibernética para minar a estabilidade política de um Estado (HATHAWAY e colab., 2012).

Porém, pode-se constatar que há congruências entre as duas definições apresentadas. Ambas afirmam que um ciberataque consiste em qualquer ação tomada com o objetivo de infligir prejuízo cibernético à parte opositora, que pode estar no nível pessoal, organizacional e até mesmo Estatal, fazendo com que sistemas e infraestruturas de rede não se comportem conforme o planejado.

Nesse sentido, pode-se afirmar que o simples uso de tecnologia cibernética para atacar cineticamente um inimigo não pode ser configurado como um ataque cibernético, mas sim como um ataque convencional tecnologicamente avançado. Por outro lado, o uso de armas cinéticas, como explosivos, focados na interrupção e/ou destruição de sistemas deve ser considerado um ataque cibernético, pois objetiva que os sistemas atingidos não se comportem como o planejado (HATHAWAY e colab., 2012).

Vale ressaltar, também, que a literatura menciona outros termos, como Guerra Cibernética e Crime Cibernético. Nesse contexto, HATHAWAY e colab. (2012) fizeram um resumo das características desses termos (Tabela 1) e seus relacionamentos (Figura 1) para diferenciar ciberataques de outras ações cibernéticas.

Figura 1 - Relação entre as ações cibernéticas



Fonte: HATHAWAY e colab., 2012

Tabela 1 - Características essenciais de ações cibernéticas

	Ataque cibernético	Crime cibernético	Guerra cibernética
Envolve somente atores não estatais		X	
Deve ser uma violação de lei criminal, cometida por meio de um computador		X	
O objetivo deve ser prejudicar o funcionamento de uma rede de computadores	X		X
Deve ter um propósito político ou de segurança nacional	X		X
Efeito deve ser equivalente a um ataque armado ou num contexto de conflito armado			X

Fonte: Adaptado de HATHAWAY e colab., 2012

No contexto deste trabalho, serão abordados somente os ataques cibernéticos que consistem em ações executadas com o uso de tecnologias cibernéticas, mais especificamente os ataques APT.

2.1 ATAQUES APT

Atualmente, governos e negócios têm enfrentado uma crescente ameaça cibernética denominada Ataque APT, que são ataques multifacetados, sofisticados, multifásicos e de longa duração focados em um alvo particular (ALGULIYEV e IMAMVERDIYEV, 2014).

Os ataques APT são capazes de se adaptar às capacidades de defesa da vítima por possuir uma forte habilidade de ocultação. O ponto de entrada, o caminho e o tempo do ataque são incertos e imprevisíveis, dificultando a detecção dele com o uso de sistemas de defesa tradicionais (LI e colab., 2016).

Outro problema relacionado aos ataques APT é a dificuldade de detecção de ataques continuados. Esses modernos ataques são extensos, temporalmente falando, e, uma vez que encontrem um ponto de entrada, ficam dormentes por longos períodos. Ademais, somente ocorre comunicação com o ambiente externo quando o ataque encontra uma oportunidade e por um curto período, tornando sua

detecção difícil, pela ausência de anormalidades óbvias na comunicação executada (LI e colab., 2016).

Os ataques APT possuem 4 características que os distinguem dos ataques tradicionais, quais sejam: alvos específicos e objetivos claros, atacantes equipados e altamente organizados, campanhas de longa duração com tentativas repetitivas e técnicas de ataque furtivas e evasivas (CHEN e colab., 2014). A seguir essas caraterísticas serão elaboradas com maior detalhe.

2.1.1 Alvos específicos e objetivos claros

Os alvos dos ataques APT são comumente governos ou organizações que possuem alto valor de propriedade intelectual. Os 10 principais alvos dos ataques em 2019 foram, respectivamente, os seguintes setores: entretenimento, finanças, governos, negócios, tecnologia, comunicações, sistema de saúde, energia, transportes e sem fins lucrativos (FIREEYE, 2020) .

2.1.2 Atacantes bem equipados e altamente organizados

Em 2019, 17 grupos APT foram considerados ativos, sendo que 6 deles são patrocinados por Estados, a saber: China, Rússia, Irã, Coreia do Norte, Vietnam e Paquistão (FIREEYE, 2020), o que comprova que diversos ataques APT são financiados com verba governamental e têm, como auxílio às suas atividades, acesso à inteligência estatal e militar (CHEN e colab., 2014).

Esses grupos são formados por militares de unidades de guerra cibernética de países ou por mercenários contratados por governos ou organizações privadas, fazendo-os bem equipados sob os pontos de vista financeiro e técnico (VIRVILIS e colab., 2013).

2.1.3 Campanhas de longa duração com tentativas repetidas

Os ataques APT são comumente longos e podem ficar sem detecção por muitos meses ou anos. Atacantes APT atacam suas vítimas de maneira persistente e buscam adaptar seus esforços quando uma tentativa prévia é frustrada.

2.1.4 Técnicas de ataque furtivas e evasivas

Uma das características que fazem os ataques APT poderem ser de longa duração é o uso de técnicas que permitem que o ataque fique dentro do tráfego de

rede da organização, interagindo somente o mínimo necessário para cumprir seus objetivos (CHEN e colab., 2014).

Como exemplo, ataques APT podem fazer uso de explorações de dia zero (*zero-day exploits*), que implica a exploração de fraquezas descobertas em softwares no mesmo dia em que ela é descoberta. Como as atualizações de segurança só serão aplicadas posteriormente, esses ataques conseguem se evadir dos sistemas de detecção de intrusão. (VIRVILIS e colab., 2013).

Resumindo, a Tabela 2 apresenta uma comparação entre ataques APT e ataques tradicionais.

Tabela 2 - Ataques tradicionais VS ataques APT

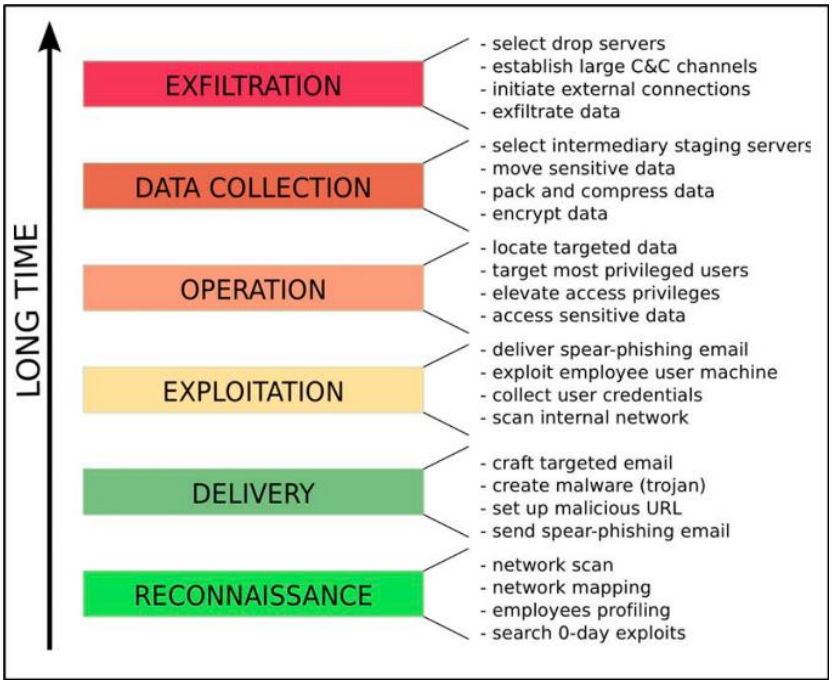
	Ataques Tradicionais	Ataques APT
Atacante	Basicamente uma única pessoa	Sofisticados e altamente organizados
Alvo	Não específico	Organizações específicas, incluindo governamentais
Propósito	Benefícios financeiros, demonstração de habilidades	Vantagens competitivas, benefícios estratégicos
Abordagem	Tentativa única, obter o que for possível num curto período de tempo	Múltiplas tentativas que ocorrem por um logo período de tempo

Fonte: adaptado de CHEN e colab., 2014

2.2 ESTÁGIOS DE UM ATAQUE APT

Os ataques APT são estruturados tipicamente em 6 estágios, quais sejam: reconhecimento, entrega, exploração, operação, obtenção de dados e exfiltração (GIURA e WANG, 2012). A Figura 2 mostra um esquema dos estágios apresentados pelos autores.

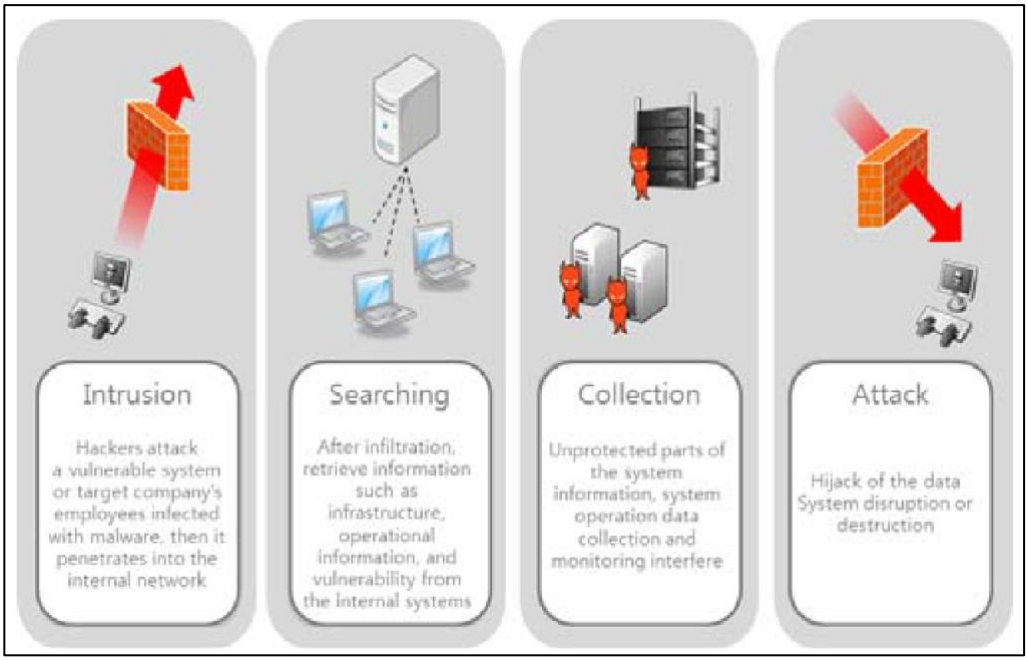
Figura 2 - Estágios de um ataque APT



Fonte: GIURA e WANG, 2012

Alguns autores estruturam um ataque APT de maneira mais simplificada. AHN e colab. (2014) agrupam os estágios de reconhecimento e entrega no estágio chamado intrusão e exploração e operação no estágio chamado busca. Assim, a estrutura de um ataque APT para esses autores pode ser evidenciada na Figura 3.

Figura 3 - Esquema simplificado de ataque APT



Fonte: AHN e colab., 2014

A estruturação de AHN e colab. (2014) mostram que o último estágio não é somente focado na exfiltração de dados, mas é também focado na interrupção de funcionamento ou destruição do sistema atacado, sendo, portanto, compatível com a definição da ciberataque que foi estudada neste capítulo.

Por outro lado, apesar de a estruturação apresentada por GIURA e WANG (2012) não se encaixar na definição de ciberataque apresentada (sendo compatível com crime cibernético), seus estágios são mais explicativos e se encaixam na estrutura de AHN e colab. (2014). Sendo assim, a estrutura daqueles autores será explicada em maior detalhe nas seções seguintes.

2.2.1 Reconhecimento

O primeiro estágio de um ataque APT é focado na coleta de informações, sendo uma importante fase de preparação para o ataque. Nesse estágio, os atacantes identificam e estudam a organização alvo. O objetivo é coletar informações sobre as tecnologias de defesa utilizadas e sobre pessoal chave para a organização. Tal estágio é, frequentemente, com o uso de engenharia social e inteligência *open-source* (CHEN e colab., 2014).

A engenharia social está diretamente relacionada com a manipulação psicológica de pessoas. Sendo assim, os atacantes podem buscar informações sobre pessoas específicas dentro da organização para criar uma reação que a faça tomar certa ação desejada. Isso pode ser caracterizado na produção de um e-mail falso com um assunto de interesse, no qual há um *link* que, se clicado, executará um malware.

Inteligência *open-source* é o conjunto de informações já previamente coletada e disponível por meio de fontes públicas. Essas informações podem ser obtidas de maneira gratuita ou paga, dependendo do perfil do alvo.

2.2.2 Entrega

Neste estágio é feita a infiltração do ataque, ou seja, o envio do código malicioso para abrir o canal de comunicação. A entrega pode ser feita de maneira direta ou de maneira indireta (CHEN e colab., 2014). A entrega direta é feita por técnicas de engenharia social, como o *spear phishing.* A entrega indireta é feita com o uso de um terceiro sujeito que é confiado pelo alvo, permitindo que a confiança nesse terceiro seja explorada, por exemplo, por meio de um *website*

frequentado por pessoas da organização, conhecido como ataque *watering hole* (GIURA e WANG, 2012).

O *spear phishing* consiste na elaboração de e-mails falsos, feitos com informações de interesse de um certo indivíduo ou grupo de pessoas. Esses e-mails falsos são elaborados como se tivessem sido enviados de fontes confiáveis, como pessoas ou organizações. Tais e-mails podem conter *links* ou anexos maliciosos, que, quando clicados, irão executar um *malware*, realizando a infiltração (GIURA e WANG, 2012).

Ataques *watering hole* começam com a identificação de *websites* comumente usados pelo pessoal da organização atacada. Com isso, os atacantes adulteram ou criam versões falsas dos *websites* selecionados para que, como decorrência da interação com esses sítios, seja possível ganhar acesso ao sistema desejado (NATARAJAN, 2017).

2.2.3 Exploração

Assim que o acesso à organização alvo é obtido, o *malware* entregue é executado, permitindo a criação de uma conexão de comando e controle (C2) entre a máquina da vítima e o atacante remoto. Isso permite que, de maneira furtiva, os atacantes continuem a obter informações referentes a configurações de seguranças, nomes de usuário e *passwords,* que poderão, também, ser úteis em ataques futuros (GIURA e WANG, 2012).

2.2.4 Operação

Assim que um caminho de entrada na rede da organização é obtido, os atacantes se mantêm na rede atacada por longos períodos de tempo. Dessa forma, é possível, por meio de movimentos horizontais na rede, identificar servidores que guardam informações sensíveis, identificar usuários que possuem as credenciais de rede necessárias para acesso mais profundo e com isso é possível criar a estratégia para os estágios de obtenção de dados e exfiltração (GIURA e WANG, 2012).

2.2.5 Obtenção de dados

Com as informações sobre servidores que possuem dados de interesse e com as credenciais de acesso obtidas nos estágios anteriores, é possível aos atacantes realizar a coleta de dados.

Nesse ponto, os dados coletados são redundados por meio de cópias em servidores internos, para garantir que mudanças de segurança não gerem perda de dados para os atacantes. Da mesma forma, os dados coletados são segmentados, comprimidos e criptografados antes da exfiltração (GIURA e WANG, 2012).

2.2.6 Exfiltração/Ataque

De acordo com CHEN e colab. (2014), a principal função de um ataque APT é roubar dados sensíveis para se obter vantagens estratégicas. Por isso, o estágio final citado por eles é chamado de exfiltração. Nesse sentido, a fase final de um ataque APT seria a transferência dos dados para servidores controlados pelos atacantes.

Cabe mencionar que a exfiltração dos dados é feita paulatinamente, para evitar a movimentação de grandes volumes de dados. Da mesma forma, os atacantes fazem uso de protocolos seguros como SSL/TLS para evitar a detecção da transmissão dos dados coletados.

Porém, cabe relembrar que um ciberataque implica a interrupção e/ou destruição do sistema atacado. Sendo assim, AHN e colab. (2014) apresentam um estágio final mais amplo, onde, além da exfiltração, o ataque APT culmina no mau funcionamento do sistema ou na sua destruição usando as credenciais obtidas nas fases anteriores.

Como forma de contextualizar, pode-se citar os ataques APT conhecido com Stuxnet1, Operation Aurora2, Operation Shady RAT3 e GhostNET[4] (VUKALOVIĆ e DELIJA, 2015).

---

[1] O *Stuxnet* é um *worm*, descoberto em 2010, que infectou sistemas de controle industrial, principalmente no Irã. O *worm* foi propagado através de *pen drives* USB infectados e explorou várias vulnerabilidades de dia zero no sistema operacional Windows. O *worm* destruiu centrífugas nucleares nas instalações de enriquecimento de urânio do Irã. Especula-se que a criação do *Stuxnet* foi um esforço conjunto entre Israel e os EUA.

[2] A operação Aurora foi uma sequência de múltiplos ataques em 2009, direcionados ao *Google*, *Adobe Systems*, *Rackspace*, *Juniper Networks* e provavelmente muitos outros. Os ataques foram originários da China e exploraram várias vulnerabilidades de dia zero no Internet Explorer. O Google teve propriedade intelectual roubada. Como resultado, o Google deixou o mercado chinês.

[3] A *Operation Shady RAT* foi uma sequência de ataques iniciados em 2006. A *McAfee*, uma empresa de cibersegurança, afirma que os ataques atingiram mais de 70 organizações, a maioria delas em os Estados Unidos. Os ataques provavelmente se originaram da China.

[4] O *GhostNET* foi uma operação de espionagem cibernética, descoberta em 2009, que tem como alvo sistemas de computador em mais de 100 países. Os invasores usaram ferramentas de *phishing* e administração remota. Os ataques são originários da China e têm como alvo governos, ministérios e embaixadas. O governo da China nega envolvimento.

3 ETAPAS DA DEFESA CIBERNÉTICA BASEADAS EM *BIG DATA*

Conforme já mencionado anteriormente, os mecanismos de defesa usuais possuem dificuldade de identificar ataques APT. Sendo assim, cada vez mais se estuda o uso do Big Data na prevenção de ciberataques.

Autores como CHEN e colab. (2014), AHN e colab. (2014), HURST e colab. (2014), MISHRA e SINGH (2016) e VIRVILIS e colab. (2013) publicaram artigos científicos que propõem o uso do *Big Data* como ferramenta de defesa cibernética, mostrando seu potencial e aplicabilidade.

Sendo assim, pode-se constatar que o uso do *Big Data* se mostra como uma fonte massiva de dados que pode ser utilizada em sistemas de prevenção de ciberataques que são executados com o uso de tecnologias cibernéticas, como ataques APT.

O uso do *Big Data* traz vantagens para a defesa cibernética. A sua capacidade de gerenciar a coleta, a consolidação e a correlação de dados de qualquer número de fontes de dados, como tráfego de rede e dispositivos de rede, fazem com que uma organização tenha uma visão holística de sua infraestrutura, permitindo correlacionar eventos de baixa severidade esporádicos com um ataque em andamento (VIRVILIS e colab., 2013).

Ademais, o *Big Data* permite a detecção de anomalias, baseado na correlação de eventos recentes e eventos históricos. Por exemplo, o aumento de tráfego DNS de um sistema em particular, por um curto período de tempo, pode ser legítimo, porém se o mesmo comportamento é identificado em uma série histórica, é possível se tratar de uma exfiltração de dados camuflada de tráfego legítimo (VIRVILIS e colab., 2013).

Sendo assim, o uso do *Big Data* permite que uma grande massa de dados possa ser analisada ao longo de períodos de tempo significantes para que a assinatura de ataques APT possa ser extraída do tráfego legítimo de rede.

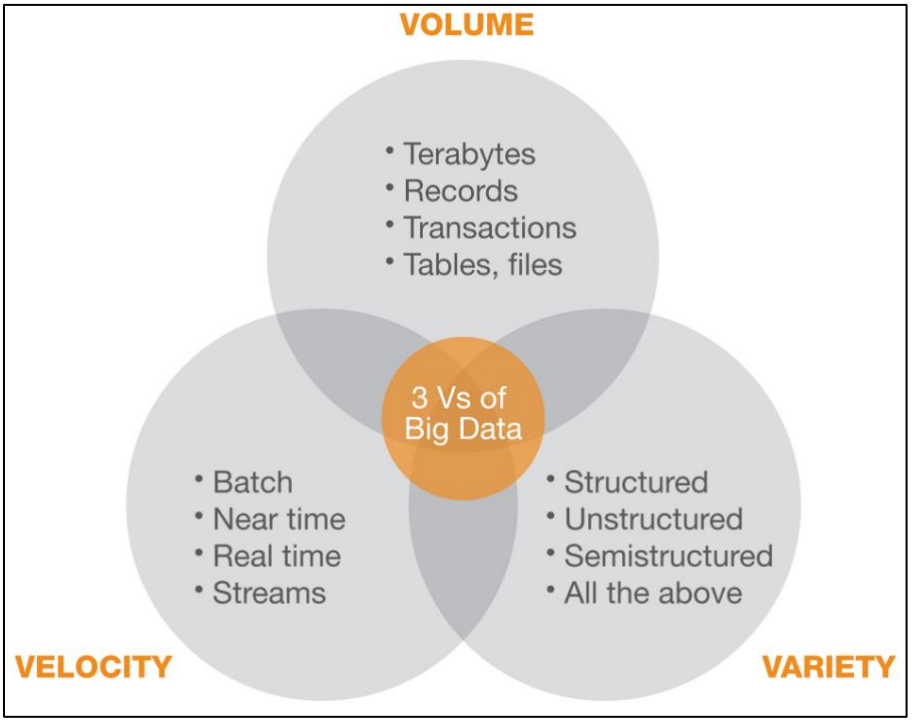
Nesse sentido, variadas arquiteturas de sistema de defesa baseados em *Big Data* estão sendo criadas e postas em testes. Diversas tecnologias e características distintas têm sido estudadas e colocadas sob testes para analisar a efetividade das soluções propostas. Porém, de modo geral, se constata que muitos sistemas possuem etapas em comum; a etapa de coleta de dados, a etapa de processamento dos dados e a etapa de análise de dados (AHN e colab., 2014), que serão abordadas em maior detalhe nos itens subsequentes.

3.1 COLETA DE DADOS

A coleta de dados é o ponto de entrada desses sistemas de defesa cibernética. Dados de diferentes fontes, e com diferentes características, são utilizados no processo de ingestão de dados em um sistema de análise baseado em *Big Data* (JI e colab., 2016).

Como já mencionado anteriormente, o *Big Data* é uma fonte massiva de dados, porém há duas outras características importantes: a variedade dos dados e a velocidade com que eles são obtidos, que em conjunto formam os 3 Vs do *Big Data* (RUSSOM, 2011), que podem ser vistos na Figura 4.

Figura 4 - Os três Vs do *Big Data*



Fonte: RUSSOM, 2011

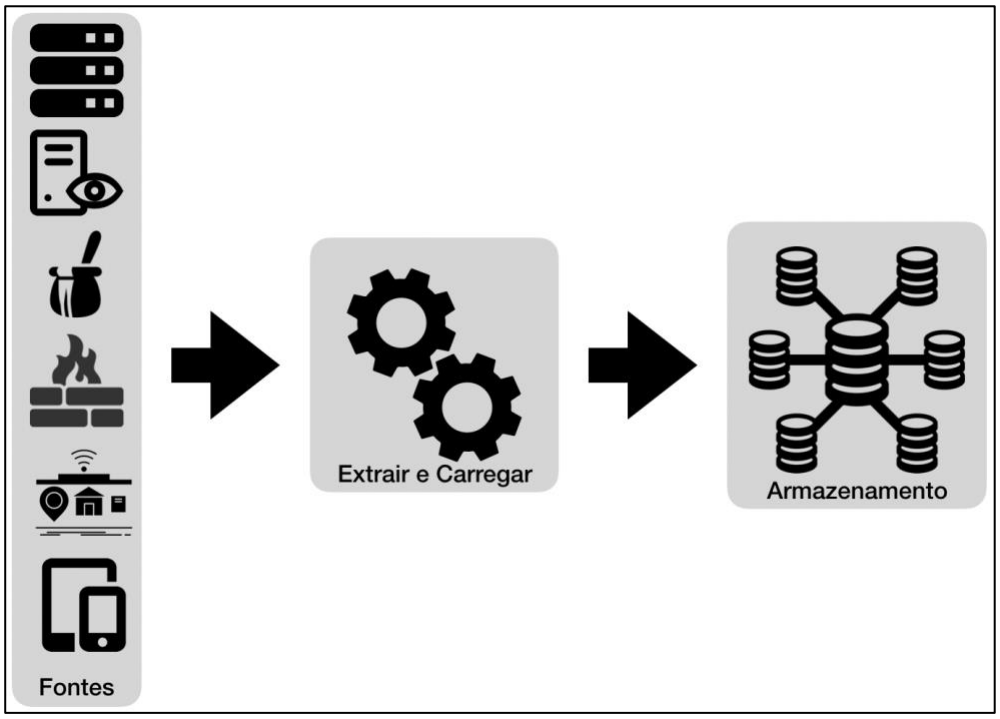
O grande volume e a velocidade com que os dados são obtidos torna necessário o uso de sistemas de armazenamento distribuídos, pois, além de permitir o acesso com maior velocidade aos dados, são escaláveis, permitindo que a capacidade de armazenamento seja ampliada com facilidade (KLEIN e colab., 2016).

A grande variedade dos dados torna necessária a utilização de ferramentas que permitam a extração e o carregamento dos dados em uma base de dados distribuída, permitindo que diferentes estruturas de dados sejam inseridas no sistema de armazenamento utilizado (KLEIN e colab., 2016).

No caso de sistemas de defesa cibernética baseados em *Big Data,* as fontes de dados são os múltiplos equipamentos que possuem conexão com a rede, como Sistemas de Detecção de Intrusão (*IDS*), *Firewalls*, dispositivos conectados, *honeypots*[5] e, até mesmo, dispositivos móveis conectados à rede são utilizados como fornecedores de informação (CAMPIOLO e colab., 2018).

Dessa forma, a coleta de dados se conecta às fontes de dados, extrai o conteúdo delas e carrega os dados num sistema de armazenamento distribuído, como se pode ver na Figura 5.

Figura 5 - Coleta de dados



Fonte: O autor

3.2 PROCESSAMENTO DE DADOS

O processamento de dados é responsável por realizar uma filtragem inicial dos dados. Isso ocorre por meio de um processo que verifica se os dados coletados atendem certos requisitos previamente estipulados. Os dados que passam pela verificação devem ser catalogados e estruturados para a etapa de análise. Vale ressaltar que, como se trabalha com um grande volume de dados, a etapa de

[5] *Honeypot* é um recurso computacional que tem a função proposital de simular falhas de segurança em um sistema para poder colher informações sobre o invasor. Funciona como uma armadilha para invasores (https://pt.wikipedia.org/wiki/Honeypot).

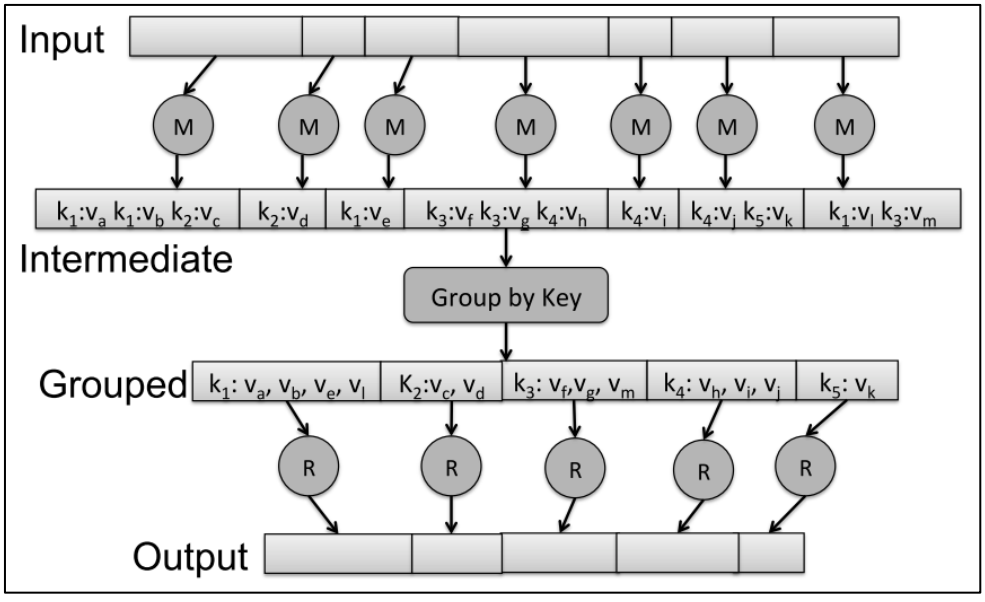
processamento se torna mais eficiente se realizada com um uso de sistemas distribuídos ou computação em nuvem, o que permite diminuir o tempo levado nesta etapa (AHN e colab., 2014).

O paradigma de processamento mais difundido no mercado é o *MapReduce* (*MR*) (RAMÍREZ-GALLEGO e colab., 2018). O *MR* foi desenvolvido para otimizar o processamento de grandes volumes de dados armazenados em sistemas distribuídos.

A principal característica do *MR* reside no uso de uma estrutura essencial: o par (chave, valor). Todo o processamento é feito com o uso de pares (chave, valor), sendo dividido em duas etapas: o mapeamento e a redução (RAMÍREZ-GALLEGO e colab., 2018).

Em linhas gerais, a etapa de mapeamento lê os dados e transforma os registros em um conjunto de pares (chave, valor). Posteriormente, a etapa de redução rearranja os dados e combina chaves coincidentes, formando novos? pares (chave, lista de valores). Finalmente, os redutores realizam uma fusão da lista de valores de cada chave, de acordo com regras previamente estipuladas, gerando o dado processado (DOULKERIDIS e NØRVÅG, 2014). A Figura 6 mostra uma concepção geral do funcionamento do *MR*.

Figura 6 - Funcionamento geral do *MapReduce*



Fonte: RAMÍREZ-GALLEGO e colab., 2018

Um exemplo costumeiramente usado para explicar o funcionamento do *MapReduce* é o processamento para se determinar o número de vezes que cada palavra aparece em um texto.

Considere o texto “No meio do caminho tinha uma pedra, tinha uma pedra no meio do caminho”. Após a coleta desse texto, ele seria dividido em partes e armazenado em um banco de dados distribuído. Como exemplo, podemos considerar sua divisão em duas partes: “No meio do caminho tinha uma pedra” e “tinha uma pedra no meio do caminho”.

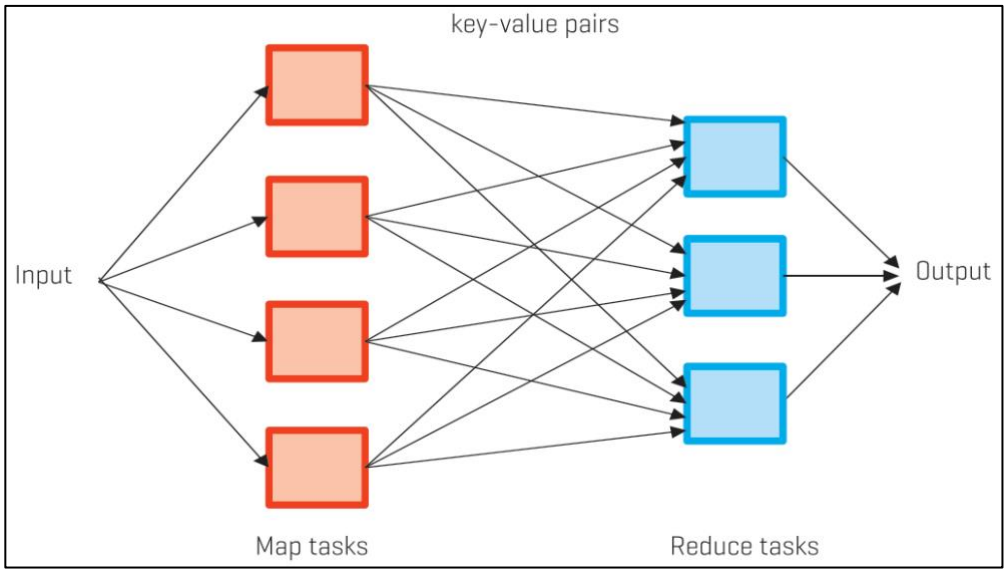
Posteriormente, seriam criados mapeadores para cada porção do texto. Esse mapeadores iriam criar pares (chave, valor), com a palavra e o número de vezes que ela aparece na porção analisada, a saber: (no, 1), (meio, 1), (do, 1), (tinha, 1), (uma, 1) e (pedra, 1), para o primeiro mapeador e (tinha, 1), (uma, 1), (pedra, 1), (no, 1), (meio, 1), (do, 1) e (caminho, 1), para o segundo mapeador.

Posteriormente, seriam criados redutores para o processamento. Esse redutores iriam pegar os resultados de cada mapeador e criar um par (chave, lista de valores), a saber: (no, [1, 1]), (meio, [1, 1]), (do, [1, 1]), (caminho, [1, 1]), (tinha, [1, 1]), (uma, [1, 1]), (pedra, [1, 1]).

Finalmente, considerando que o objetivo é contar a quantidade de vezes que cada palavra aparece, a fusão feita na redução seria executada por meio de uma soma, gerando o seguinte resultado final (no, 2), (meio, 2), (do, 2), (caminho, 2), (tinha, 2), (uma, 2), (pedra, 2), no qual se constata que cada palavra aparece duas vezes no texto analisado.

Todo o processamento do algoritmo ocorre em paralelo. A tarefa é dividida entre múltiplos mapeadores e múltiplos redutores. Essa abordagem permite que a carga de trabalho seja dividida entre os servidores que compõem a estrutura distribuída, fazendo que com cada máquina faça apenas uma porção do trabalho, conforme se observa na Figura 7.

Figura 7 - Paralelismo do MapReduce



Fonte: ULLMAN, 2012

Ao fim do processamento, o dado terá sido estruturado para posterior uso na etapa de análise, a qual será a responsável final para transformar os dados armazenados em informações de interesse. Ou seja, a etapa de análise irá criar conhecimento.

No caso de sistemas de defesa cibernética baseados em *Big Data*, alguns autores propõem que o processamento pode ser feito com base no comportamento dos *hosts* da rede protegida. Esse comportamento pode ser obtido diariamente por meio da coleta das seguintes características: número de *bytes* transmitidos por cada *host* para endereços externos, número de conexões de cada *host* para *hosts* externos e número de *IPs* externos relacionados com uma conexão iniciada pelo *host* interno. Assim, é possível, na etapa de análise, identificar se *hosts,* em um dado momento, estão envolvidos em atividades de rede suspeitas, possivelmente relacionadas à fase de exfiltração de dados de um ataque APT (MARCHETTI e colab., 2016).

Por outro lado, outros autores propõem um processamento baseado nos processos de um *host*, com o objetivo de identificar quais deles estão se comportando de maneira anormal. Nesse contexto, KIM e colab. (2014) apresentam as características da Tabela 3, que permitem a criação de pares (chave, valor), onde a chave é o identificador de um dado processo e o valor é uma

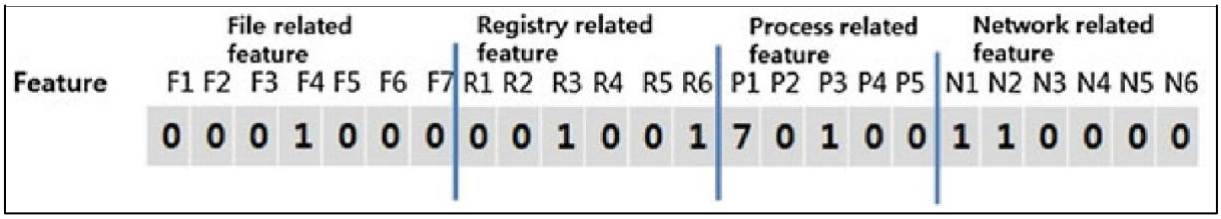
estrutura montada com o número de vezes que cada característica é identificada (Figura 8).

Tabela 3 - Características de processo

Tipo	Característica	Código
Arquivo	Exclusão de arquivo na pasta do sistema	F1
	Renomeação de arquivo na pasta do sistema	F2
	Criação de arquivo na pasta do sistema	F3
	Criação de arquivo na pasta temporária	F4
	Criação de arquivo executável	F5
	Criação de arquivo na pasta temporária	F6
	Criação de arquivo	F7
Registro	Exclusão do *registry*	R1
	Exclusão de serviço	R2
	Adicionando execução automática	R3
	Registro de *registry*	R4
	Registro de serviço	R5
	Adicionando um item BHO	R6
Processo	Criação de outro processo	P1
	Encerramento de outro processo	P2
	Pesquisa de outro processo	P3
	Injeção de código DLL	P4
	Criação de *thread*	P5
Rede	Abertura de porta	N1
	Ligação de porta	N2
	Conexão de rede	N3
	Desconexão da rede	N4
	Envio de dados	N5
	Recebimento de dados	N6

Fonte: Adaptado de KIM e colab., 2014

Figura 8 - Estrutura das características de processo



Fonte: KIM e colab., 2014

Vale ressaltar que há muitas outras características presentes em eventos de rede que podem ser utilizadas durante a etapa de processamento, para posterior aproveitamento na etapa de análise. Como exemplo, pode-se citar as 41 características presentes na base de dados NSL-KDD, que é a base de dados padrão para pesquisas relacionadas à detecção de intrusão de sistemas (AGGARWAL e SHARMA, 2015).

3.3 ANÁLISE DE DADOS

A etapa de análise utiliza os dados estruturados da fase de processamento e, com o auxílio de algoritmos de predição, classificação e associação, permite identificar se uma dada atividade de rede é potencialmente danosa. Os algoritmos de predição mais comuns são baseados em regressões que analisam o passado e o presente para prever uma tendência futura. Os algoritmos de associação agrupam atividades de rede similares e os algoritmos de classificação identificam atividades de rede em classes, notadamente atividade normal ou ataque (AHN e colab., 2014).

A análise por meio de aprendizado de máquina permite que os algoritmos sejam treinados para identificar automaticamente atividades de rede que sejam ameaças, sem a necessidade da intervenção humana. Desta forma, resta ao tomador de decisão atuar com base no que for identificado pelo sistema de defesa cibernética. Vale ressaltar que toda a análise dos dados pode se valer do uso de Inteligência Artificial (IA) (ALVES, 2018), porém este trabalho está limitado a técnicas de classificação por meio de aprendizado de máquina.

Durante a presente pesquisa, verificou-se que todos os autores estudados trabalharam com técnicas de classificação. Da mesma forma, foi constatado que as técnicas mais comumente utilizadas são *Support Vector Machines* (*SVM*), Árvore de Decisão, *Naive Bayes, K-Nearest Neighbors* (*KNN*) e *K-Means*, o que

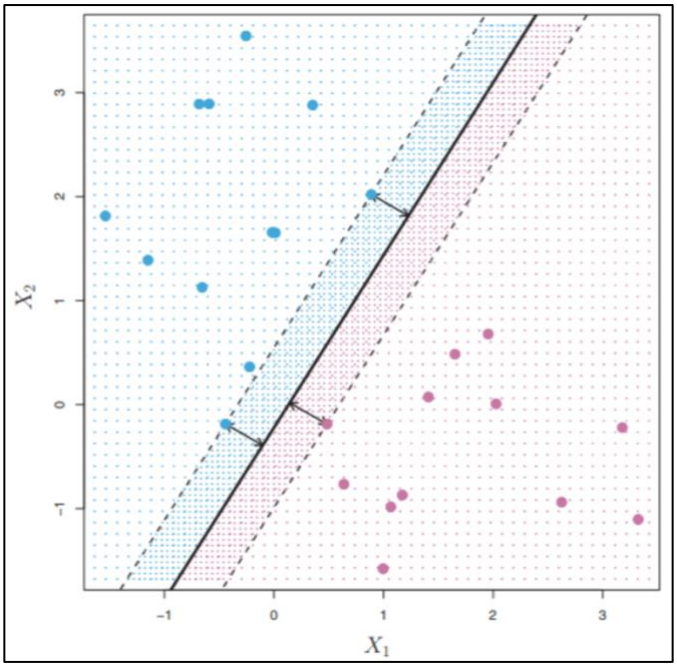
está de acordo com o citado por XIN e colab. (2018), limitando as técnicas apresentadas aqui.

3.3.1 Support Vector Machines (SVM)

SVM é um dos métodos mais robusto e acurado em todos os algoritmos de aprendizado de máquina (XIN e colab., 2018). A ideia principal por trás desse método é buscar o melhor hiperplano que separa um conjunto de dados. Assim, todos os dados que caírem de um lado do hiperplano terão uma classificação associada e os dados que caírem do outro lado do hiperplano terão uma classificação oposta associada, o que é muito útil na determinação de atividades de rede normais ou anormais (CHU e colab., 2019).

No caso de dados representados em 1D, o hyperplano é um ponto. No caso de dados representados em 2D, o hyperplano é uma linha. No caso de dados representados em 3D, o hyperplano é um plano. Dessa forma, pode-se concluir que caso os dados sejam representados em um espaço n-dimensional, o hyperplano que o SVM determinará será representado em um espaço (n-1)-dimensional (SIROHI, 2019).

Figura 9 - Dados 2D com o respectivo hyperplano 1D



Fonte: SIROHI, 2019

As características do algoritmo o fazem ser considerado rápido (BUCZAK e GUVEN, 2016) e sua alta acurácia (CHU e colab., 2019) o fazem ser largamente

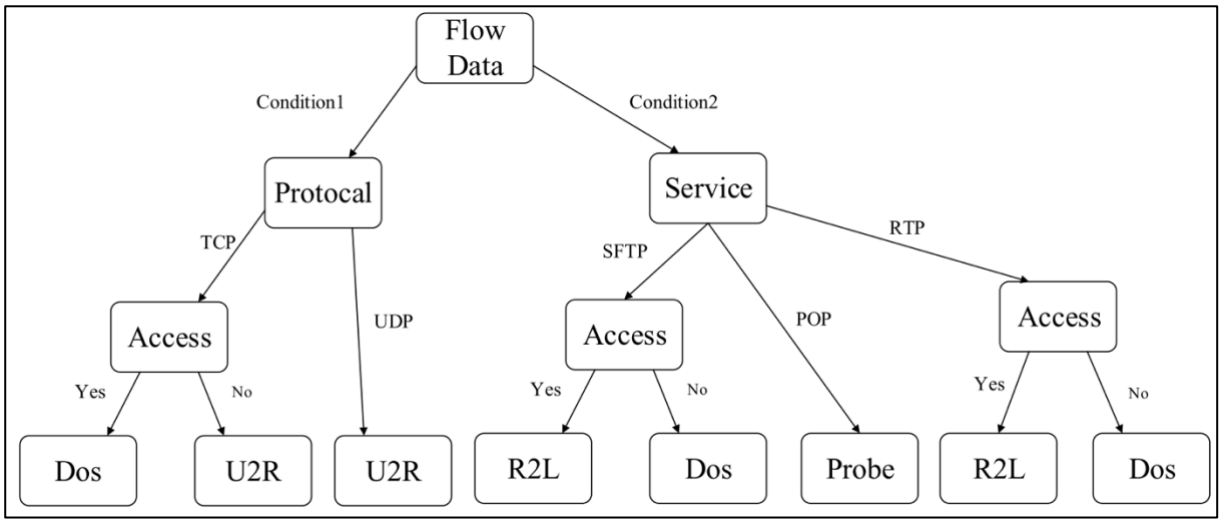
utilizado, sendo utilizado em vários trabalhos científicos, como apresentado por (XIN e colab., 2018).

3.3.2 Árvore de decisão

O algoritmo da árvore de decisão é formado com o uso dos dados utilizados para o treinamento. Sua estrutura é formada por uma raiz, nós e folhas. O primeiro nível da árvore é a raiz, sendo um nó particular que não possui nenhum nível acima. Os demais nós, possuem níveis acima e abaixo. Por fim, as folhas são o fim da árvore, somente possuindo nós acima (BUCZAK e GUVEN, 2016).

Cada nó da árvore representa um teste feito em uma das características selecionadas nos dados de treinamento. A seleção de qual característica deve ser utilizada em cada nó é feita por meio de algoritmos, dos quais pode-se citar o ID3, o C4.5 e o CART como os mais famosos (XIN e colab., 2018). A Figura 10 mostra um exemplo simplificado de árvore de decisão montada com algumas características que podem ser utilizadas para classificar eventos de rede.

Figura 10 - Exemplo de árvore de decisão



Fonte: XIN e colab., 2018

Com a árvore montada, um evento qualquer, representado por um vetor de características, pode ser classificado percorrendo a árvore a partir do nó raiz. A cada nó, deve se verificar o teste condicional, seguindo para o próximo nó até que se chegue em uma das folhas da árvore.

Uma das vantagens desse método recai em sua simplicidade de implementação. BUCZAK e GUVEN (2016) afirmam que algoritmos como o C4.5

são mais simples que algoritmos complexos como os utilizados no SVM, o que torna o uso desse método interessante.

3.3.3 Naive bayes

O classificador *Naive Bayes* pode lidar com um número arbitrário de características. Apesar de ter diversas limitações, ele é um classificador ótimo se as características são independentes. Comumente, ele é um dos primeiros classificadores a ser comparado com classificadores mais complexos como o SVM e uma de suas maiores vantagens é a possibilidade de seu treinamento ser realizado em tempo linear (BUCZAK e GUVEN, 2016).

Naive Bayes prevê o resultado da classificação de acordo com o teorema Bayesiano. Ou seja, a determinação da classe de um dado evento (vetor de x características) é feita por meio do cálculo das probabilidades condicionais de cada classe dado o evento analisado, a saber: $P(Ci|x1, x2, \dots, xn)$.

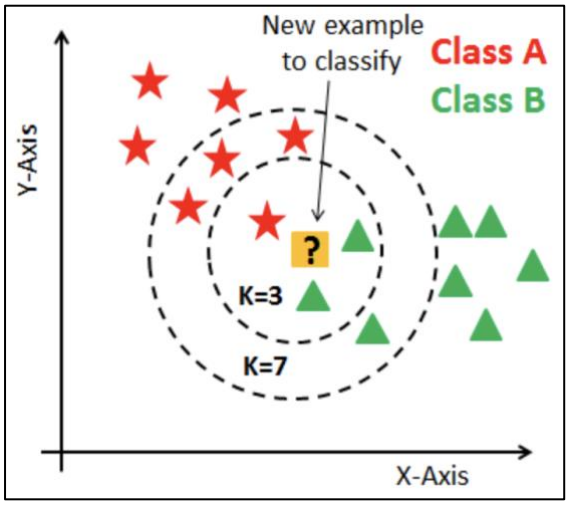
A equação $P(Ci|x1, x2, \dots, xn) = \frac{P(Ci)[P(x1|Ci)P(x2|Ci)\dots P(xn|Ci)]}{P(x1)P(x2)\dots P(xn)}$ é obtida do teorema de Bayes e que pode ser resumida para $P(Ci|X) = \frac{P(Ci)P(X|Ci)}{P(X)}$, o que significa dizer que a probabilidade de Ci dado o evento X é proporcional a probabilidade de Ci vezes a probabilidade de X dado Ci (CHU e colab., 2019).

As probabilidades condicionais $P(x1|Ci), P(x2|Ci), \dots, P(xn|Ci)$ e as probabilidades $P(Ci)$ de cada classe são calculadas durante a etapa de treinamento (BENFERHAT e colab., 2008). Posteriormente, durante a fase de classificação, para cada classe é feito o cálculo de probabilidade condicional $P(Ci|X)$ e a classe com maior valor é atribuída ao evento X (CHU e colab., 2019).

3.3.4 K-Nearest Neighbors (KNN)

O classificador *KNN* é baseado em uma função distância que mede diferença ou similaridade entre duas instâncias. Essa função distância é calculada por meio da distância euclidiana padrão. Nesse método de classificação, os dados de treinamento são utilizados como conjunto de vizinhos, para que, em seguida, sejam determinados os k vizinhos mais próximos a uma dada entrada. Por fim, a moda das classes dos k vizinhos é atribuída como classe à entrada, finalizando a classificação, o que torna o knn um algoritmo de simples implementação (BUCZAK e GUVEN, 2016).

Figura 11 - Algoritmo KNN



Fonte: NAVLANI, 2018

No caso de sistemas de defesa cibernética, os dados de treinamento são um conjunto de vetores de características, como as presentes no banco de dados NSL-KDD (AGGARWAL e SHARMA, 2015), já agrupadas em classes. O dado de entrada, utilizado para o cálculo dos k-vizinhos mais próximos, é um vetor com mesmas características que se deseja classificar de acordo com sua similaridade com seus vizinhos.

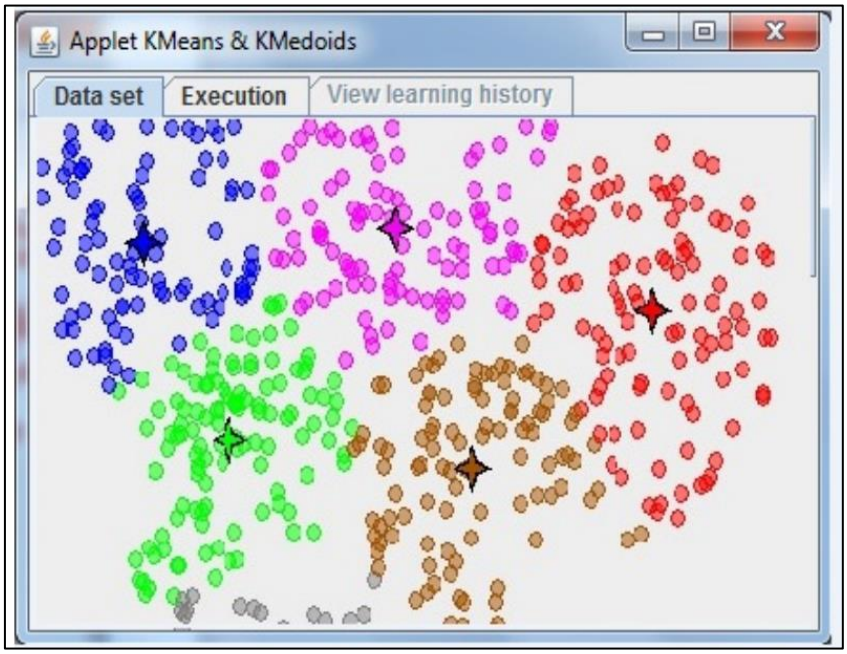
3.3.5 K-Means

O K-Means é um método de partição utilizado para criar agrupamentos de observações similares. Seu método de agrupamento analisa observações (vetores de características) e determina a qual grupo elas devem pertencer por meio de cálculos de proximidade (RAZAQ e colab., 2016).

Inicialmente, um conjunto de dados de treinamento é utilizado para criar os grupos. O primeiro passo é determinar, de maneira manual ou automática, quantos grupos deverão ser criados (valor k). Posteriormente, de maneira aleatória, se seleciona k observações de treinamento para serem os centroides dos grupos. Em seguida, as demais observações são agrupadas de acordo com a proximidade do centroide mais próximo. Após a determinação dos grupos, é feito o cálculo atualizado dos centroides e todo o processo de agrupamento é refeito. Quando não

houver mais alterações significativas nos valores dos centroides, o algoritmo é parado (ARORA e colab., 2016).

Figura 12 - Algoritmo K-Means



Fonte: (ARORA e colab., 2016)

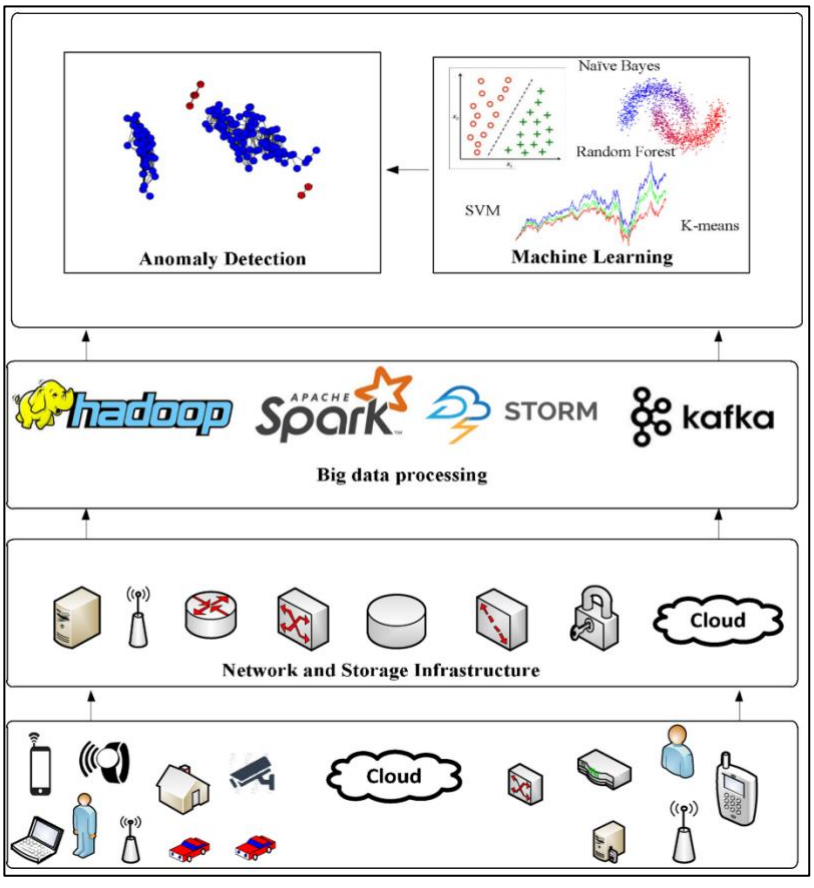
4 ARQUITETURAS ESTUDADAS

Este capítulo tem por objetivo apresentar algumas arquiteturas de sistemas de defesa cibernética baseada em *Big Data* que foram estudadas durante esta pesquisa. Até o momento, foram estudadas as definições de ataque cibernético – com detalhamento dos ataques APT – e foram caracterizadas as etapas principais desses sistemas. Com isso, é possível ter os principais conceitos envolvidos nessa complexa área de conhecimento.

De posse desses conceitos, é possível analisar implementações reais de tais sistemas, com o objetivo de entender e usá-los como base (o que mais for aplicável) para desenvolver ou evoluir um sistema de defesa cibernética baseada em *Big Data* para o EB.

Nas arquiteturas estudadas, pode-se observar, além da presença das etapas apresentadas no Capítulo ETAPAS DA DEFESA CIBERNÉTICA BASEADAS EM *BIG DATA*, o compartilhamento de tecnologias utilizadas. Dessa forma, a Figura 13 mostra a concepção geral da arquitetura desses sistemas.

Figura 13 - Concepção geral das arquiteturas



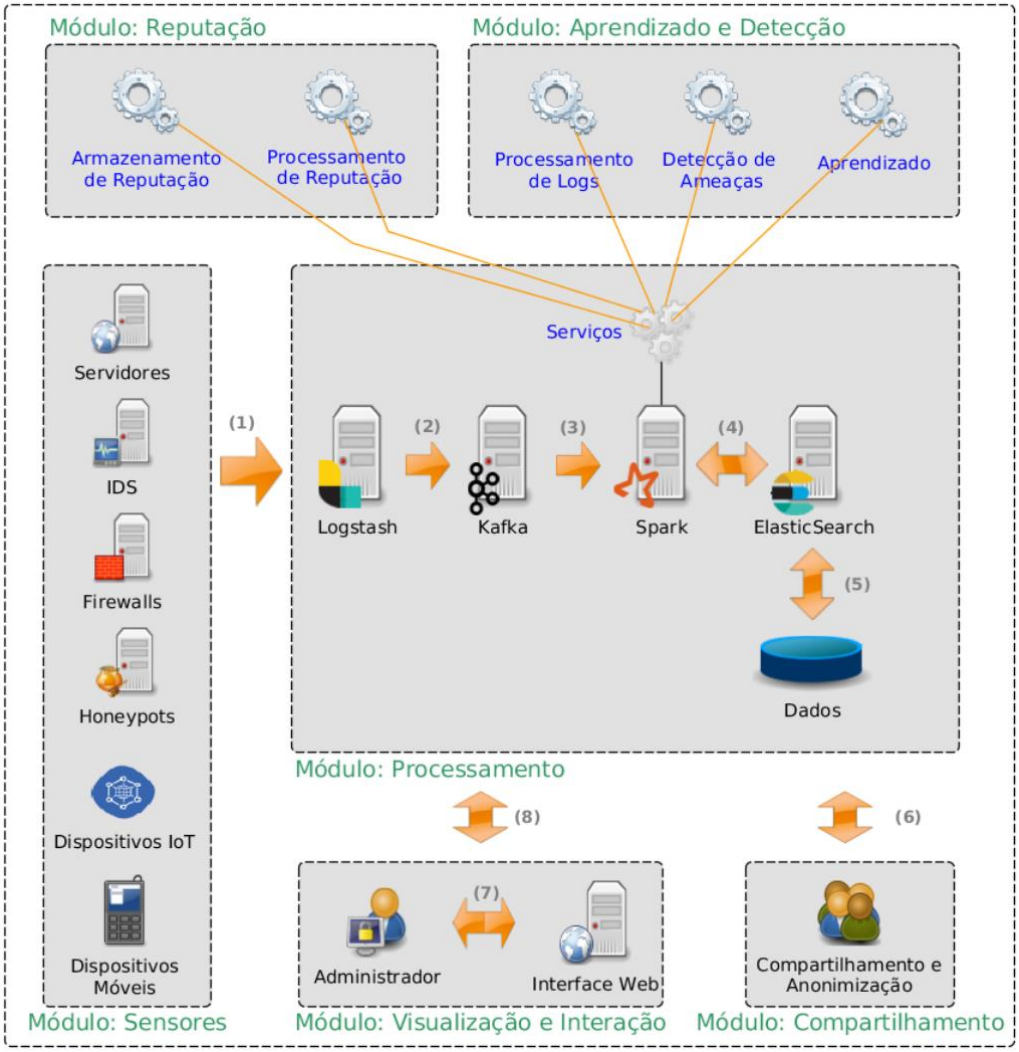
Fonte: (ARIYALURAN HABEEB e colab., 2019)

Nesse sentido, nas subseções seguintes, serão apresentadas algumas das arquiteturas estudadas, com detalhamento das tecnologias utilizadas, para posterior discussão.

4.1 PROPOSTA DE CAMPIOLO E COLAB. (2018)

A arquitetura apresentada por CAMPIOLO e colab. (2018) pode ser vista na Figura 14. Nela, podemos ver todas as etapas da defesa cibernética baseada em *Big Data* apresentadas no presente trabalho.

Figura 14 - Arquitetura de Campiolo e colab.



Fonte: CAMPIOLO e colab., 2018

Nessa arquitetura, pode-se observar o uso de tecnologias como *Logstash, Kafka, Spark e ElasticSearch*, que serão detalhadas a seguir.

4.1.1 Logstash

O *Logstash*[6] é uma ferramenta de coleta de dados que pode, dinamicamente, unir dados de diferentes fontes e normalizar os dados coletados em diferentes destinos (CAMPIOLO e colab., 2018). Sendo assim, o *Logstash* está intimamente ligado com a etapa de coleta de dados (seção 3.1), mais precisamente à extração de dados das mais diferentes fontes, como mostrado na Figura 5.

4.1.2 Kafka

O *Apache Kafka*[7] é, atualmente, o *framework* mais famoso para realizar a ingestão de dados de forma distribuída, para possibilitar o processamento paralelo de um grande volume de dados (LE NOAC'H e colab., 2017). Sendo assim, assim como o *Logstash,* está intimamente ligado com a etapa de coleta de dados, mais precisamente ao carregamento dos dados em bancos de dados distribuídos.

4.1.3 Spark

O *Framework Apache Spark*[8] surgiu em 2010. Ele provê suporte específico para diversas atividades relacionadas ao uso de *Big Data*, dentre os quais pode-se citar seu motor de processamento e biblioteca de aprendizado de máquina (RAMÍREZ-GALLEGO e colab., 2018).

Spark é considerado a plataforma de *Big Data* em tempo-real mais popular e tem por objetivo fazer a etapa de análise de dados ser executada com mais rapidez que outras soluções, oferecendo a capacidade de processar grandes volumes de dados em memória (JI e colab., 2016).

Dessa forma, as características do *Spark,* principalmente a capacidade de processar dados distribuídos e analisar dados por meio de algoritmos de aprendizado de máquina, o fazem estar intimamente relacionado com as etapas de processamento (seção 3.2) e análise (seção 3.3).

4.1.4 Elasticsearch

O *ElasticSearch*[9] gerencia grandes volumes de dados, desde o armazenamento até a recuperação. Ele foi desenvolvido, inicialmente, como um

[6] https://www.elastic.co/pt/logstash
[7] https://kafka.apache.org/
[8] https://spark.apache.org/
[9] https://www.elastic.co/

sistema de pesquisa textual em grandes volumes de dados não estruturados. Atualmente, o *ElasticSearch* é um sistema de análise de dados com várias capacidades (VOIT e colab., 2017). Dessa forma, esse mecanismo está intimamente ligado com a etapa de análise (seção 3.3).

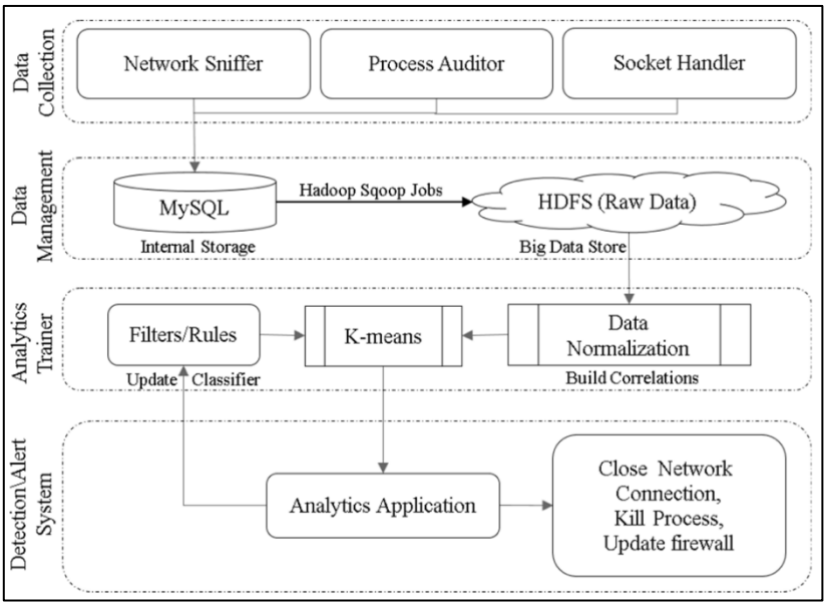
4.1.5 Funcionamento

De modo resumido, o *Logstash* coleta dados dos sensores e o *kafka* faz o carregamento deles em um banco de dados distribuído, o que permite o processamento em paralelo. Posteriormente, o *Spark* faz o processamento e a análise dos dados. Por fim, as análises executadas são armazenadas por meio do uso do *ElasticSearch* (CAMPIOLO e colab., 2018).

4.2 PROPOSTA DE RAZAQ E COLAB. (2016)

A arquitetura apresentada por RAZAQ e colab. (2016) pode ser observada na Figura 15. Nela, podemos ver todas as etapas da defesa cibernética baseada em *Big Data* apresentadas no presente trabalho.

Figura 15 - Arquitetura de Razaq e colaboradores



Fonte: RAZAQ e colab., 2016

Nessa arquitetura pode-se notar o uso de tecnologias como MySql, Hadoop Sqoop e HDFS (componente *Data Management*). No componente *Analysis Trainer*, observa-se o uso do *K-Means* (seção 3.3.5) como algoritmo de classificação.

4.2.1 MySQL

O *MySQL*[10] é um sistema gerenciador de bancos de dados (sgbd) que é largamente utilizando e, além de ter um custo de total de propriedade baixo (TCO[11]), pode ser utlizado em muitas plataformas (*Windows*, *Linux*, *macOS* e etc) e é estável (STEVE SUEHRING, 2002, p. 43). Dessa forma, o *MySQL* está ligado à etapa de coleta de dados.

4.2.2 Hadoop Sqoop

O *Hadoop Sqoop* é um *framework* do ecossistema *Haddop*[12] que permite a transferência de dados de vários sgbd para o HDFS. Ele também permite a transferência de dados entre o HDFS e outros sgbd. Dessa forma, o *Haddop Sqoop* esta ligado à etapa de coleta de dados.

4.2.3 HDFS

O *Hadoop Distritbuted File System* (HDFS) é um sistema de dados distribuído tolerante a falhas. O HDFS é um módulo do *Hadoop* e permite o armazenamento de um grande volume de dados e é capaz de lidar com falhas em importantes partes do armazenamento sem perder dados (KUMAR e colab., 2018). O HDFS é altamente escalável, permitindo a introdução de novas máquinas para armazenamento de acordo com a necessidade (MAHMOOD e AFZAL, 2013). Ele funciona dividindo os dados em blocos, armazenando-os de maneira distribuída e com replicação, tornando-o tolerante a falhas (KUMAR e colab., 2018). Dessa forma, o HDFS esta ligado à etapa de coleta de dados.

4.2.4 Funcionamento

De modo resumido, o sistema funciona coletando dados de diferentes fontes, armazenando-os temporariamente no *MySQL,* até que os dados sejam definitivamente transferidos para o HDFS. Posteriormente, os dados armazenados no HDFS são processados, criando vetores de características úteis para identificar anomalias. Por fim, os dados processados são analisados por meio do classificador *K-Means* para se identificar observações anômalas que podem potencialmente ser ataques cibernéticos genuínos (RAZAQ e colab., 2016).

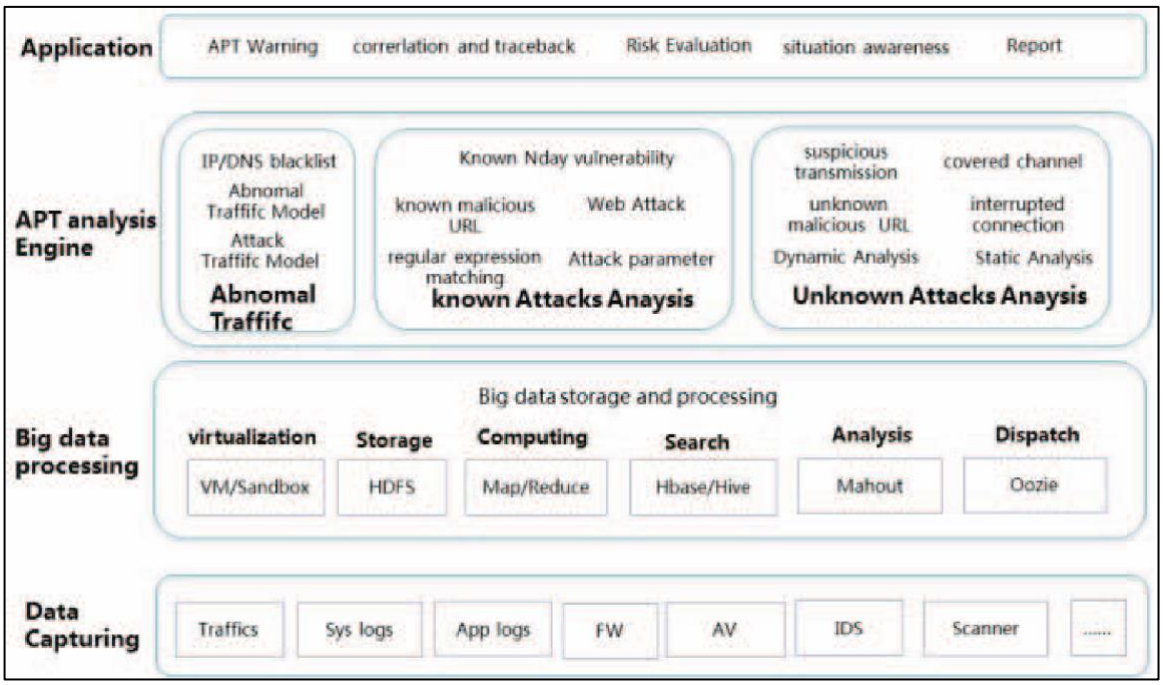
---

[10] https://www.mysql.com
[11] https://pt.wikipedia.org/wiki/Total_cost_of_ownership
[12] http://hadoop.apache.org

4.3 PROSPOSTA DE SHENWEN E COLAB. (2015)

A arquitetura proposta por SHENWEN e colab. (2015) pode ser vista na Figura 16. Nela, podemos ver todas as etapas da defesa cibernética baseada em *Big Data* apresentadas no presente trabalho.

Figura 16 - Arquitetura proposta por Shenwen e colaboradores



Fonte: SHENWEN e colab., 2015

Nessa arquitetura é possível observar que, da mesma forma como as outras apresentadas, as fontes de dados são variadas. Ademais, no componente *Big data processing*, verifica-se tecnologias já apresentadas, como *HDFS* (seção 4.2.3), *MapReduce* (seção 3.2) e outras como *HBase, Hive, Mahout e Oozie*.

4.3.1 Hbase

O *HBase*[13] faz parte do ecossistema *Haddop*. Ele é um banco de dados *NoSQL*[14] que se posiciona acima do *HDFS*, permitindo que operações sejam executadas em tempo real. O HBase tem grande utilidade para se trabalhar com dados não estruturados e é utilizado para armazenamento e processamento de dados (NEWMAN, 2019), o que o faz ter relação com as etapas de processamento de dados.

[13] https://hbase.apache.org
[14] https://pt.wikipedia.org/wiki/NoSQL

4.3.2 Hive

O Hive[15] também faz parte do ecossistema *Hadoop.* Ele é um sistema de armazenamento de dados que permite consultar grandes volumes de dados não estruturados. O *Hive* pode consultar dados armazenados no *HDFS* ou até mesmo dados armazenados no *HBase*, por meio do *MapReduce* (seção 3.2) ou *Spark* (seção 4.1.3). A principal característica do *Hive* é permitir a utilização de recursos *SQL*[16] para consultar os dados (NEWMAN, 2019). Dessa forma, o Hive tem ligação com a etapa de processamento de dados.

4.3.3 Mahout

O *Mahout*[17] também faz parte do ecossistema *Hadoop.* Ele é um motor de análises que provê algoritmos de aprendizado de máquina especificamente para análise de ataques APT, tais como classificação, agrupamento, recomendação de filtragem, mineração de padrões frequentes e etc. Dessa forma, o *Mahout* está intimamente ligado à etapa de análise de dados (SHENWEN e colab., 2015).

4.3.4 Oozie

O *Ozzie*[18] também faz parte do ecossistema *Hadoop*. Ele é um sistema de gerenciamento e agendamento de fluxos de trabalho para o *Hadoop*. Ele permite o agendamento de tarefas variadas que podem ser executas automaticamente (SHENWEN e colab., 2015). Como o ecossistema do *Hadoop* é variado e complexo, montado com o uso de diversas ferramentas e linguagens, o Oozie provê um framework para gerenciar efetivamente toda essa variedade, permitindo um controle unificado e o processamento paralelo de múltiplos trabalhos (ISLAM e colab., 2012).

4.3.5 Funcionamento

De modo resumido, assim como os outros, o sistema de defesa funciona coletando dados de múltiplas fontes. Posteriormente, esses dados são armazenados com o uso o HDFS. O processamento é feito por meio do *MapReduce,* sendo que as consultas aos dados que serão processados são feitas

[15] https://hive.apache.org
[16] https://pt.wikipedia.org/wiki/SQL
[17] https://mahout.apache.org
[18] https://oozie.apache.org

por meio do *HBase* ou *Hive.* Por fim, a análise de tráfego anormal de rede, ataques conhecidos e ataques desconhecidos é feita por meio do *Mahout,* com o uso do algoritmo KNN (seção 3.3.4). Cabe ressaltar que todas as atividades de coleta, processamento e análise são gerenciadas pelo *Oozie*.

5 DISCUSSÃO

Em todas as arquiteturas apresentadas pode-se observar a existência das 3 principais etapas executadas por sistemas de defesa cibernética baseados em *Big Data,* a saber: coleta de dados, processamento de dados e análise de dados.

Apesar de todos os sistemas terem essas mesmas etapas, em termos de tecnologias utilizadas, verifica-se que os sistemas adotam soluções variadas para executar cada uma delas. A grande disponibilidade de tecnologias permite diferentes formas de arquitetar um sistema de defesa cibernética baseado em *Big Data*, o que pode tornar difícil escolher a melhor forma de implementar esses sistemas.

De modo geral, por serem de código aberto, pode-se observar uma predominância de soluções feitas com o uso do ecossistema *Hadoop* e outros produtos livres desenvolvidos pela *Apache*[19] (*Spark* e *Kafka*). O ecossistema *Hadoop* é uma plataforma de software que permite o processamento de grandes volumes de dados de maneira distribuída. Ele foi desenvolvido com o conceito do *MapReduce* da *Google,* no qual os dados são quebrados em blocos para então serem processados. Atualmente, o Hadoop possui diversos componentes, tais como o *MapReduce, HDFS, HBase, Hive, OOzie* (KUMAR e colab., 2018), todos apresentados neste trabalho.

Mas assim como o *Hadoop*, há diversas outras tecnologias (algumas mencionadas neste trabalho), comerciais ou não, que podem ser utilizadas e combinadas de diferentes formas para atingir um mesmo objetivo. Sendo assim, surge o questionamento de como escolher as tecnologias e os detalhes da arquitetura de um sistema de defesa cibernética baseada em *Big Data.* Para resolver esse questionamento, KLEIN e colab. (2016) propuseram uma arquitetura de referência para sistemas baseados em *Big Data* no domínio da segurança nacional, o que tem grande potencial de uso para o Exército Brasileiro em suas atividades de defesa cibernética.

A arquitetura de referência apresentada pelos autores é focada em necessidade típicas no domínio da defesa nacional e é tecnologicamente neutra, ou seja, não está amarrada em nenhuma tecnologia. Segundo os autores, há diversas arquiteturas de referência publicadas na comunidade científica, mas elas

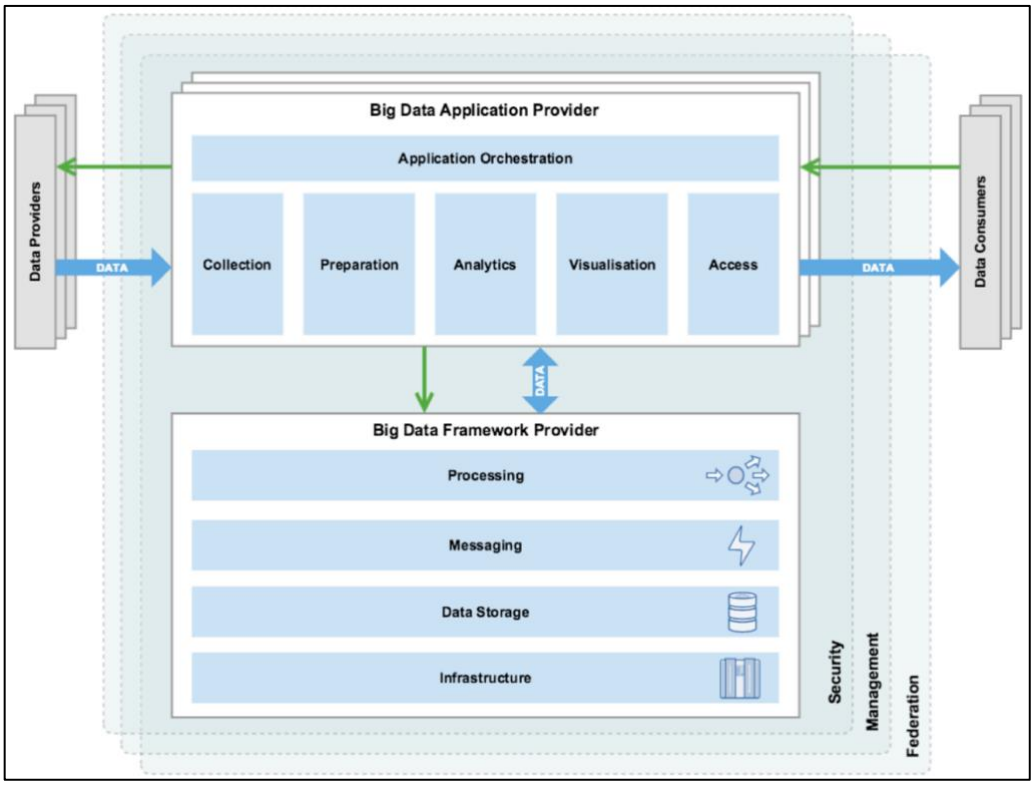
[19] https://www.apache.org

não são normalmente aplicáveis aos clientes no domínio da segurança nacional, porque são muito genéricas ou porque são muito amarradas em tecnologias específicas.

A arquitetura proposta por eles é dividida em três grandes componentes, como pode-se constatar na Figura 17. O Provedor de Aplicação de *Big Data* inclui a lógica de negócio do nível da aplicação, as transformações de dados e análise, e funcionalidade para ser executada pelo sistema. O Provedor de Framework de *Big Data* inclui *software,* armazenamento, plataformas computacionais e redes usadas pelo Provedor de Aplicação de *Big Data.* A Figura 17 também mostra que é possível haver várias instâncias do Provedor de Aplicação de *Big Data.*

O terceiro componente são os módulos transversais, que abordam preocupações referentes à segurança e ao gerenciamento do sistema, e à sua interoperabilidade com outros sistemas nacionais.

Figura 17 - Arquitetura de referência



Fonte: KLEIN e colab., 2016

Outro importante passo citado pelos autores é mapear as tecnologias disponíveis, tanto as comerciais como as de código aberto, para que seja possível atender as necessidades dos patrocinadores e usuários do sistema que será implementado. Os autores informam que mapearam 35 produtos para o Provedor

de Aplicação de *Big Data* e 64 produtos para o Provedor de Framework de *Big Data*. A grande quantidade de produtos disponíveis mostra que esse mapeamento é vital para decidir o rumo tecnológico da implementação de um sistema de defesa cibernética baseada em *Big Data*.

Por fim, os autores apresentam uma lista de perguntas que devem ser respondidas para guiar o desenvolvimento de um sistema de defesa cibernética baseada em *Big Data*, a saber:

1. Visualização – Que informação os usuários precisam?
2. Coleta – Quais são as fontes de dados e como coletar os dados?
3. Análise – Que informação precisa ser extraída dos dados?
4. Preparação – Quais preparações são necessárias antes da análise?
5. Armazenamento – Como os dados serão armazenados para suportar análise, visualização e acesso?
6. Processamento – Como as análises serão executadas?
7. Orquestração da Aplicação – O fluxo de processamento precisa de uma orquestração?
8. Acesso – Qual API de acesso é requerida? Como os dados serão acessados?
9. Mensagens – É necessária uma infraestrutura de mensagens de suporte?
10. Gerenciamento – Como a aplicação e a infraestrutura serão gerenciadas?
11. Segurança – Quais controles de segurança são necessários?
12. Federação – A solução precisa se interligar com outros sistemas nacionais?
13. Infraestrutura – Qual infraestrutura é necessária?

Com a posse das respostas das perguntas, é possível analisar as tecnologias mapeadas para se determinar quais delas se adaptam mais às características desejadas. Sendo assim, acredita-se que, apesar da multiplicidade de tecnologias de *Big Data,* é possível utilizar a arquitetura de referência e a metodologia apresentada para se desenvolver um sistema de defesa cibernética baseado em *Big Data* que possa ser implementado pelo Exército Brasileiro.

6 CONCLUSÃO

Do estudo das definições dadas pelo Governo Americano e pela OCX, pode-se observar que um ciberataque consiste em qualquer ação tomada com o objetivo de infligir prejuízo cibernético à parte opositora, fazendo com que sistemas e infraestruturas de rede não se comportem conforme o planejado. Isso permite concluir que um sistema de defesa cibernético eficiente deve ser abrangente o suficiente para impedir ou mitigar ameaças nos níveis pessoal, organizacional e até mesmo Estatal.

Ademais, na atualidade, constata-se que governos têm sido alvos frequentes de ciberataques, estando entre os 10 principais alvos no ano de 2019. Sendo assim, considerando que o Exército Brasileiro, por meio de seu Objetivo Estratégico número 4 (OEE 4), deve ser capaz de atuar no espaço cibernético com liberdade de ação, conclui-se que esse assunto possui grande relevância para a Força Terrestre (FTer).

Nesse contexto, a multiplicidade de sistemas de defesa cibernética baseados em *Big Data* tem mostrado o seu grande potencial de uso, principalmente no que se refere aos sofisticados ataques APT. Tais sistemas são costumeiramente projetados de maneira modular, o que permite que diferentes tecnologias possam ser utilizadas no desenvolvimento, garantindo flexibilidade no produto final.

Posto isso, como forma de identificar tecnologias, foram apresentados alguns sistemas de defesa cibernética baseados em *Big Data.* Da análise dos mesmos, pôde-se verificar quais tecnologias foram utilizadas na implementação desses sistemas, sendo possível observar uma prevalência de tecnologias livres desenvolvidas pela *Apache* (*Spark, Kafka* e ecossistema *Hadoop*), o que permite concluir que é possível desenvolver um sistema de defesa cibernética sem arcar com custos de licenciamento de *software* especializado.

Finalmente, foi verificado que, apesar de todos os sistemas contarem com as mesmas etapas principais, há uma grande variedade de tecnologias e de detalhes de arquitetura que permitem a implementação de diferentes soluções para criar um sistema de defesa cibernética baseada em *Big Data*. Dessa forma, objetivando fornecer um direcionamento, foi apresentada uma arquitetura de referência focada nas necessidades típicas no domínio da defesa nacional que pode ser utilizada como base no desenvolvimento de sistemas de defesa cibernética baseados em *Big Data* pelo Exército Brasileiro.

Assim, como conclusão deste trabalho, pode-se afirmar que o uso do *Big Data* em sistemas de defesa cibernética não somente é viável, como é necessário para prevenir e detectar ataques cibernéticos sofisticados como os ataques APT. Portanto, constata-se que o desenvolvimento de sistemas de defesa cibernética baseados em *Big Data* deve ser buscado ativamente pelo Exército Brasileiro.

Por fim, cabe ressaltar que este trabalho não visa esgotar o assunto e, devido à sua atualidade e relevância, sugere-se, como linha de pesquisa para trabalhos futuros, a continuidade de pesquisas ligadas ao uso do *Big Data* em sistemas de defesa cibernética.

REFERÊNCIAS

AGGARWAL, Preeti e SHARMA, Sudhir Kumar. **Analysis of KDD Dataset Attributes - Class wise for Intrusion Detection**. Procedia Computer Science, v. 57, p. 842–851, 2015. Disponível em: <http://dx.doi.org/10.1016/j.procs.2015.07.490>.

AHN, Sung Hwan e KIM, Nam Uk e CHUNG, Tai Myoung. **Big data analysis system concept for detecting unknown attacks**. International Conference on Advanced Communication Technology, ICACT, p. 269–272, 2014.

ALGULIYEV, Rasim e IMAMVERDIYEV, Yadigar. **Big Data: Big promises for information security**. 8th IEEE International Conference on Application of Information and Communication Technologies, AICT 2014 - Conference Proceedings, 2014.

ALVES, P M M R. **O Impacto De Big Data Na Atividade De Inteligência**. Revista Brasileira de Inteligência, v. 1, n. 13, p. 1–20, 2018. Disponível em: <http://www.abin.gov.br/conteudo/uploads/2018/12/RBI-13_artigo-2_O-IMPACTO-DE-BIG-DATA-NA-ATIVIDADE-DE-INTELIGÊNCIA.pdf>.

ARIYALURAN HABEEB, Riyaz Ahamed e colab. **Real-time big data processing for anomaly detection: A Survey**. International Journal of Information Management, v. 45, n. August, p. 289–307, 2019. Disponível em: <https://doi.org/10.1016/j.ijinfomgt.2018.08.006>.

ARORA, Preeti e DEEPALI e VARSHNEY, Shipra. **Analysis of K-Means and K-Medoids Algorithm for Big Data**. Physics Procedia, v. 78, n. December 2015, p. 507–512, 2016. Disponível em: <http://dx.doi.org/10.1016/j.procs.2016.02.095>.

BAILES, Alyson J. K. e colab. **The shanghai cooperation organization**. Stockholm International Peace Research Institute, n. 17, p. 60, 2007.

BENFERHAT, Salem e KENAZA, Tayeb e MOKHTARI, Aicha. **A Naive Bayes approach for detecting coordinated attacks**. Proceedings - International Computer Software and Applications Conference, p. 704–709, 2008.

BRASIL. **Política Nacional de Inteligência**. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2016/Decreto/D8793.htm>. Acesso em: 1 mar 2020.

BUCZAK, Anna L. e GUVEN, Erhan. **A Survey of Data Mining and Machine**

Learning Methods for Cyber Security Intrusion Detection. IEEE Communications Surveys and Tutorials, v. 18, n. 2, p. 1153–1176, 2016.

CAMPIOLO, Rodrigo e colab. **Uma Arquitetura para Detecção de Meaças Ciberneticas Baseada na Analise de Grandes Volumes de Dados**. Anais do I Workshop de Segurança Cibernética em Dispositivos Conectados, p. 1–5, 2018. Disponível em: <https://ojs.sbc.org.br/index.php/wscdc/article/view/2401>.

CHEN, Ping e DESMET, Lieven e HUYGENS, Christophe. **A study on advanced persistent threats**. IFIP International Conference on Communications and Multimedia Security, p. 63–72, 2014.

CHU, Wen Lin e LIN, Chih Jer e CHANG, Ke Neng. **Detection and classification of advanced persistent threats and attacks using the support vector machine**. Applied Sciences (Switzerland), v. 9, n. 21, 2019.

DOULKERIDIS, Christos e NØRVÅG, Kjetil. **A survey of large-scale analytical query processing in MapReduce**. VLDB Journal, v. 23, n. 3, p. 355–380, 2014.

FIREEYE. **M-Trends 2020**. Disponível em: <https://content.fireeye.com/m-trends/rpt-m-trends-2020>. Acesso em: 10 mar 2020.

GALLAHER, Ryan. **Facing data deluge, secret U.K. spying report warned of intelligence failure**. Disponível em: <https://theintercept.com/2016/06/07/mi5-gchq-digint-surveillance-data-deluge/>. Acesso em: 28 fev 2020.

GIURA, Paul e WANG, Wei. **A context-based detection framework for advanced persistent threats**. Proceedings of the 2012 ASE International Conference on Cyber Security, CyberSecurity 2012, n. SocialInformatics, p. 69–74, 2012.

HATHAWAY, Oona A. e colab. **The law of cyber-attack**. California Law Review, v. 100, n. 4, p. 817–885, 2012.

HURST, William e MERABTI, Madjid e FERGUS, Paul. **Big data analysis techniques for cyber-threat detection in critical infrastructures**. Proceedings - 2014 IEEE 28th International Conference on Advanced Information Networking and Applications Workshops, IEEE WAINA 2014, p. 916–921, 2014.

ISLAM, Mohammad e colab. **Oozie: Towards a scalable workflow management system for hadoop**. ACM International Conference Proceeding Series, 2012.

JAMES E., Cartwrigth. **Memorandum for Chiefs of the Military Services Commanders of the Combatant Commands, Directors of the Joint Staff Directorates on Joint Terminology for Cyberspace Operations**. Disponível em: <http://www.nsci-va.org/CyberReferenceLib/2010-11-joint Terminology for Cyberspace Operations.pdf>. Acesso em: 5 mar 2020.

JI, Cun e colab. **IBDP: An Industrial Big Data Ingestion and Analysis Platform and Case Studies**. Proceedings - 2015 International Conference on Identification, Information, and Knowledge in the Internet of Things, IIKI 2015, p. 223–228, 2016.

KIM, Hyunjoo e KIM, Ikkyun e CHUNG, Tai-Myoung. **Frontier and innovation in future computing and communications**. Lecture Notes in Electrical Engineering, v. 301, p. 553–563, 2014.

KLEIN, John e colab. **A Reference Architecture for Big Data Systems in the National Security Domain**. 2nd International Workshop on BIG Data Software Engineering, p. 51–57, 2016. Disponível em: <https://sci-hub.tw/https://ieeexplore.ieee.org/abstract/document/7811387/metrics>.

KUMAR, Praveen e colab. **Analysis and comparative exploration of elastic search, MongoDB and Hadoop big data processing**. Advances in Intelligent Systems and Computing, v. 584, p. 605–615, 2018.

LE NOAC'H, Paul e COSTAN, Alexandru e BOUGÉ, Luc. **A performance evaluation of Apache Kafka in support of big data streaming applications**. Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017, v. 2018- Janua, p. 4803–4806, 2017.

LI, Meicong e colab. **The study of APT attack stage model**. 2016 IEEE/ACIS 15th International Conference on Computer and Information Science, ICIS 2016 - Proceedings, 2016.

LIU, Donglan e colab. **Research and application of APT attack defense and detection technology based on big data technology**. ICEIEC 2019 - Proceedings of 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication, n. 52062617002, p. 701–704, 2019.

MAHMOOD, Tariq e AFZAL, Uzma. **Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools**. Conference Proceedings

- 2013 2nd National Conference on Information Assurance, NCIA 2013, p. 129–134, 2013.

MARCHETTI, Mirco e colab. **Analysis of high volumes of network traffic for Advanced Persistent Threat detection**. Computer Networks, v. 109, p. 127–141, 2016. Disponível em: <http://dx.doi.org/10.1016/j.comnet.2016.05.018>.

MISHRA, Aditya Dev e SINGH, Youddha Beer. **Big data analytics for security intelligence**. International Conference on Computing, Communication and Automation, p. 50–53, 2016.

NATARAJAN, Krithika. **A Study On WHA (Watering Hole Attack)–The Most Dangerous Threat To The Organisation**. International Journal of Innovations in Scientific and Engineering Research (IJISER), v. 4, n. 8, p. 196–198, 2017.

NAVLANI, Avinash. **KNN Classification using Scikit-learn**. Disponível em: <https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn>. Acesso em: 11 abr 2020.

NEWMAN, Saggi. **Hive vs. HBase**.

OCX. **Agreement on Cooperation in Ensuring International Information Security between the Member States of the Shanghai Cooperation Organization**. Disponível em: <http://eng.sectsco.org/load/207508/>. Acesso em: 5 mar 2020.

RAMÍREZ-GALLEGO, Sergio e colab. **Big Data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce**. Information Fusion, v. 42, p. 51–61, 2018.

RAZAQ, Abdul e TIANFIELD, Huaglory e BARRIE, Peter. **A big data analytics based approach to anomaly detection**. Proceedings - 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, BDCAT 2016, p. 187–193, 2016.

RUSSOM, Philip. **BIG DATA ANALYTICS**. TDWI best practices report, fourth quarter, v. 19, n. 4, p. 1–34, 2011. Disponível em: <https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf>.

SHENWEN, Lin e YINGBO, Li e XIONGJIE, Du. **Study and research of APT detection technology based on big data processing architecture**. ICEIEC 2015 - Proceedings of 2015 IEEE 5th International Conference on Electronics Information and

Emergency Communication, n. 2012, p. 313–316, 2015.

SIROHI, Kshitiz. **Support Vector Machine (Detailed Explanation)**. Disponível em: <https://towardsdatascience.com/support-vector-machine-support-vector-classifier-maximal-margin-classifier-22648a38ad9c>. Acesso em: 11 abr 2020.

STEVE SUEHRING. **MySQL Bible**. New York: Wiley Publishing, Inc., 2002.

ULLMAN, Jeffrey D. **Designing good MapReduce algorithms**. XRDS: Crossroads, The ACM Magazine for Students, v. 19, n. 1, p. 30, 2012.

VIRVILIS, Nikos e SERRANO, Oscar e DANDURAND, Luc. **Big Data Analytics for Sophisticated Attack Detection**. Cis.Aueb.Gr, v. 3, p. 1–8, 2013. Disponível em: <http://www.cis.aueb.gr/Publications/ISACA - Big data analytics for intrusion detection.pdf>.

VOIT, Aleksei e colab. **Big Data Processing for Full-Text Search and Visualization with Elasticsearch**. International Journal of Advanced Computer Science and Applications, v. 8, n. 12, p. 76–83, 2017.

VUKALOVIĆ, J. e DELIJA, D. **Advanced Persistent Threats - Detection and defense**. 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015 - Proceedings, n. May, p. 1324–1330, 2015.

XIN, Yang e colab. **Machine Learning and Deep Learning Methods for Cybersecurity**. IEEE Access, v. 6, n. c, p. 35365–35381, 2018.

ZUECH, Richard e KHOSHGOFTAAR, Taghi M. e WALD, Randall. **Intrusion detection and Big Heterogeneous Data: a Survey**. Journal of Big Data, v. 2, n. 1, 2015.