UNIVERSIDADE FEDERAL FLUMINENSE
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE TELECOMUNICAÇÕES

PEDRO SILVEIRA PISA

Identificação e Previsão de Padrões Recorrentes de Distribuição de Permissões no Controle de Acesso a Serviços de Computação em Nuvem

NITERÓI
2024

UNIVERSIDADE FEDERAL FLUMINENSE
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE TELECOMUNICAÇÕES

PEDRO SILVEIRA PISA

Identificação e Previsão de Padrões Recorrentes de Distribuição de Permissões no Controle de Acesso a Serviços de Computação em Nuvem

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e de Telecomunicações da Universidade Federal Fluminense, como requisito parcial para obtenção do título de Doutor em Engenharia Elétrica e de Telecomunicações. Área de concentração: Sistemas de Telecomunicações.

Orientador:
Diogo Menezes Ferrazani Mattos

NITERÓI
2024

(Espaço reservado para a ficha catalográfica)

PEDRO SILVEIRA PISA

Identificação e Previsão de Padrões Recorrentes de Distribuição de Permissões no Controle de Acesso a Serviços de Computação em Nuvem

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e de Telecomunicações da Universidade Federal Fluminense, como requisito parcial para obtenção do título de Doutor em Engenharia Elétrica e de Telecomunicações. Área de concentração: Sistemas de Telecomunicações.

BANCA EXAMINADORA

_____

Prof. Diogo Menezes Ferrazani Mattos, D.Sc. – Orientador, UFF

_____

Prof. Dianne Scherly Varela de Medeiros, D.Sc. – UFF

_____

Prof. Rodrigo de Souza Couto, D.Sc. – UFRJ

_____

Prof. Igor Monteiro Moraes, D.Sc. – UFF

_____

Prof. Pedro Braconnot Velloso, Dr. – UFRJ

Niterói

Fevereiro de 2024

À minha família, à Julia e à Ana Rosa.

Agradecimentos

Agradeço, principalmente, à minha família e à minha companheira Julia Miceli e nossa filha Ana Rosa, que me apoiam e me acompanham em todas as horas e desafios, sempre trazendo uma palavra de força e esperança, confiando sempre no meu trabalho e nos resultados. Agradeço também ao meu Orientador Diogo Menezes, pela amizade e por todas as conversas, orientações e direções para a concretização deste trabalho, e o alcance dos objetivos propostos. Obrigado também a todos os colegas do MídiaCom e dos projetos de pesquisa que atuei, pelas ideias e sugestões, que certamente enriqueceram este trabalho. Agradeço ao meu sócio e amigo Filipe Barretto, por todos os anos de parceria, confiança, cumplicidade e amizade. Sem ela não conseguiríamos ter força e coragem para continuar vencendo nossos desafios diários e construir uma equipe forte como a que formamos na Solvimm e, agora, na e-Core. Agradeço a todos os colaboradores da antiga Solvimm e da e-Core pelo apoio, incentivo e aprendizado nas discussões de tecnologia que levaram a muitos dos elementos deste trabalho, em especial, para a Jéssica Alcântara, que semanalmente, focou em me apoiar nas discussões das propostas deste trabalho, com contribuição em muitas das conclusões aqui apresentadas.

Resumo

A tecnologia de computação em nuvem revolucionou o mercado de construção e operação de serviços de tecnologia. Com essa facilidade, diversas empresas criaram suas aplicações sem ter conhecimento adequado de segurança, o que tem ocasionado cada vez mais vazamentos de dados e invasões nas aplicações em nuvem. Esses vazamentos ocorrem por conta de configurações de acesso muito permissivas feitas por profissional sem o devido conhecimento de premissas de segurança e privilégio mínimo, que são adotadas quando o profissional foca apenas em facilidade de uso. Além do baixo conhecimento nas possibilidades de configuração, o fato do acesso aos serviços hospedados na nuvem ser obrigatoriamente remota, via estruturas existentes de conexão, há o desafio de gerenciamento de uso de rede para os administradores de redes, em especial aquelas de larga escala como a de uma Universidade. Este trabalho tem por objetivo propor uma análise de propostas e construção de uma base de dados para análise de privilégios no uso dos ambientes em nuvem e um outro algoritmo que classifique e segregue o tráfego de rede em ambientes de alta escala, uma vez que os serviços em nuvem fazem com que o perfil de uso das redes seja uma agregação dos tráfegos internos para a conexão com a Internet. Concluiu-se que através da estratégia proposta é possível estimar a carga de uso em cada ponto de acesso da rede por perfil, possibilitando a construção de soluções de detecção de anomalia de redes, planejamento e economia de energia através do desligamento de dispositivos ou ainda na implantação de soluções inteligentes de cache de dados e controle de qualidade de serviço.

Palavras-chaves: computação em nuvem, segurança, controle de acesso, privilégio mínimo, gerenciamento de rede, qualidade de serviço

Abstract

Cloud computing technology has revolutionized the market for building and operating technology applications. With this innovation, several companies created their applications without having adequate knowledge of security, which has caused more and more data leaks and intrusions into cloud applications. These leaks occur due to very permissive access settings made by a professional without proper knowledge of security and least privilege assumptions, which are adopted when the professional focuses only on ease of use. In addition to the low knowledge of configuration possibilities, the fact that the access to services hosted in the cloud is obligatorily remote, via existing network connection structures, there is the challenge in the management of network usage for network administrators, especially the ones with large scale such as that of a University. The objective of this work is to analyze proposals and build a dataset for privileges analysis in the use of cloud environments and another algorithm that classifies and segregates network traffic in high-scale environments, since cloud services make the network usage profile an aggregation of internal traffic for the Internet connection. It was concluded that through the proposed strategy it is possible to estimate the usage load in each access point of the network.

Keywords: cloud computing, security, access control, least privilege, network management, quality of service.

Lista de Figuras

Sumário

Capítulo 1 - Introdução

A tecnologia de computação em nuvem revolucionou o mercado de construção e operação de serviços de tecnologia da informação, dando acesso a um ambiente de alta qualidade, com maior oferta de serviços e conformidade a regulamentos de segurança, e estabilidade a preços mais baratos do que nas soluções de centros de dados tradicionais [1]. Essa revolução possibilitou inúmeros novos negócios e viabilizou a inovação em milhões de empresas, visto que o custo das falhas de experimentos foi reduzido a frações do que era no ambiente de servidores tradicionais. Além disso, a disponibilidade de recursos e serviços de tecnologia da informação fez com que os desenvolvedores e *startups* pudessem publicar suas aplicações mais rapidamente.

Essa disponibilidade, por outro lado, trouxe um desafio para as empresas e grandes organizações. Como os principais serviços utilizados por uma organização são hospedados fora da rede local, torna-se necessário que todos os colaboradores, usuários e visitantes da organização acessem serviços através da conexão com a Internet [2]. Isso exerce especial pressão sobre a infraestrutura de rede, em especial quando se trata de estruturas de larga escala, como redes de acesso sem fio em Universidades. Este trabalho, foca em desafios criados pela larga adoção da computação em nuvem e que não apresentam soluções comerciais viáveis.

Outro desafio criado pelo paradigma de computação em nuvem reside no permissionamento das suas funcionalidades. Com a facilidade de uso da computação em nuvem, diversos usuários da nuvem criaram suas aplicações sem ter conhecimentos adequados de segurança, como o correto controle de acesso e permissionamento no ambiente de infraestrutura [3]. Isso pode ser observado pelos anúncios, cada vez mais frequentes, de vazamentos de dados [4] e de invasões nos ambientes de nuvem, que ocorrem principalmente por erros de configuração e extravio de credenciais de acesso ao ambiente de nuvem [5].

Os objetivos deste trabalho são: análise do tráfego de rede de alta capacidade da Universidade Federal Fluminense (UFF), determinando os perfis de tráfego em sua rede, e análise de propostas e construção de uma base de dados com registros de atividades (*logs*) de ambientes de produção de clientes de um provedor de nuvem para análise de privilégio no uso dos ambientes em nuvem.

O presente trabalho está estruturado sob a forma de dois estudos independentes, expostos de forma a manter a coerência com os objetivos propostos. São apresentadas duas propostas originais, cada qual, com introdução específica relacionada ao seu objetivo. Sendo o primeiro estudo refere-se à análise do tráfego de rede de alta capacidade da Universidade Federal Fluminense (UFF), determinando os perfis de tráfego na rede da Universidade e o sugestionamento para gerenciar melhor a utilização dos diversos pontos de acesso. E o segundo estudo, uma proposta que consiste em uma análise e construção de base de dados para algoritmo gerador de políticas de acessos para garantir o privilégio mínimo no uso dos ambientes em nuvem. Essa proposta se diferencia das analisadas nos trabalhos relacionados por correlacionar políticas com os serviços em utilização e comparar esses dados entre clientes distintos da nuvem. A coleta de dados para a análise do algoritmo se dá a partir de ambientes de produção na nuvem da AWS com a consolidação em um ambiente de um repositório de dados (*datalake*) do projeto.

O trabalho está estruturado em 4 Capítulos, sendo o Capítulo 2 destinado a análise do primeiro estudo deste trabalho. Neste capítulo, apresenta-se também a análise matemática e experimental da proposta. No Capítulo 3, analisamos o segundo estudo, trazemos seus principais desafios de segurança, o estado da arte para garantir o privilégio mínimo para usuários e aplicações em ambientes de computação em nuvem é analisado e as vantagens e desvantagens de cada abordagem são discutidas. Por fim, o Capítulo 7 apresenta as conclusões.

Capítulo 2 - Estudo 1 - Análise do Tráfego de Rede de Acesso de Grande Escala

Um desafio que as empresas encontram com a adoção da computação em nuvem é o perfil de uso de rede dentro das suas redes de acesso. Nos ambientes tradicionais, os sistemas internos da empresa são acessados através de redes locais e privadas, enquanto com a computação em nuvem, os sistemas são acessados através da Internet. Isso faz com que os sistemas internos exijam maior proteção de rede e o gerenciamento do tráfego de rede, tanto dentro da rede quanto na borda, seja aperfeiçoado para identificar e priorizar os fluxos de rede relativos aos sistemas da empresa. Neste Capítulo, é abordado a primeira proposta trabalhada, que envolve o gerenciamento de tráfego de rede de acesso de larga escala, como a de uma Universidade.

As redes de acesso sem fio se caracterizam pela crescente convergência de diversos serviços. Logo, há elevada demanda de disponibilização de espectro de frequências de *Wi-Fi* para atender aos usuários simultâneos que buscam altas taxas de transmissão. Assim, a predição de carga em cada ponto de acesso é essencial para alocar recursos e auxiliar o complexo projeto de dimensionamento da rede. Contudo, a carga em cada ponto de acesso varia com o número de dispositivos conectados e com as características do tráfego de cada dispositivo em dado instante. Este Capítulo propõe uma estratégia baseada em cadeias de Markov para a previsão do número de dispositivos conectados aos pontos de acesso da rede sem fio e aplica um modelo de aprendizado de máquina não supervisionado para a identificação dos perfis de tráfego. Os principais objetivos são identificar padrões de tráfego e pontos de sobrecarga na rede sem fio e, então, dimensionar eficientemente a rede e prover uma base de conhecimento para ferramentas de segurança. A proposta é avaliada na rede de grande escala da Universidade Federal Fluminense, com 670 pontos de acesso distribuídos em uma ampla área. Os dados coletados são anonimizados e o processamento ocorre na nuvem. Os resultados mostram que a proposta é capaz de prever o número de dispositivos conectados com 90% de precisão e agrupa cinco perfis de tráfego que definem na rede sem fio.

As redes móveis sem fio são cada vez mais ubíquas e apresentam crescimento constante e significativo em sua adoção. As estruturas dessas redes cresceram 71% em 2017 e ao fim desse mesmo ano o volume de dados transferidos por mês alcançou 11,5 milhões de *exabytes*. O crescimento acumulado é de 17 vezes nos últimos 5 anos [6], sendo

fortemente impulsionado pela larga adoção de tecnologias 4G e 5G no mundo. No entanto, existe uma tendência global de migração dos dados para redes sem fio estruturadas, variantes do padrão IEEE 802.11 (*Wi-Fi*). A carga nessas redes deve ultrapassar os 100 *exabytes* por mês em 2022, tornando as redes *Wi-Fi* responsáveis por mais de 59% do tráfego móvel.

As redes móveis foram inicialmente usadas para troca de mensagens, consultas rápidas e acessos a sites. Atualmente, são largamente utilizadas para aplicações multimídia, como chamadas de voz, redes sociais e acesso a vídeos, englobando diversos perfis de uso. Estima-se que em 2022, 79% do tráfego das redes sem-fio será de vídeo [6], exigindo uma adaptação da rede para essa nova demanda. O aumento na carga de redes *Wi-Fi* exige o consequente aumento da capacidade e disponibilidade da rede. É necessário, portanto, dimensioná-las adequadamente para atender à quantidade de usuários e seus diversos perfis de uso. No entanto, dimensionar a rede e prever o tráfego demandado são tarefas desafiadoras, pois a carga sobre os pontos de acesso é um processo estocástico e a grande quantidade de pontos de acesso em operação contribui para a existência de interferência constante entre os pontos de acesso vizinhos [7]. Além disso, uma grande oferta de pontos de acesso leva o cliente a trocar frequentemente o ponto de acesso ao qual está associado, pois a potência percebida na recepção pelo cliente apresenta alta variabilidade [8]. Esse cenário fomenta a criação de mecanismos inteligentes de análise do tráfego de rede que sejam cientes do perfil de acesso esperado dos usuários. Ademais, os mecanismos inteligentes devem fornecer dados concretos para dimensionar adequadamente a rede, além de bases para a detectar usos anômalos da rede.

Este Capítulo propõe uma estratégia para prever a carga nos pontos de acesso sem fio e aplica um modelo de aprendizado de máquina não supervisionado para identificar perfis de tráfego dos usuários. A carga em um ponto de acesso é representada pelo número de dispositivos conectados a ele. O objetivo é obter informações que permitam dimensionar a rede adequadamente, de acordo com o perfil de uso da rede em cada localidade e para cada usuário. A estratégia proposta se baseia em uma arquitetura para captura de dados nas redes monitoradas e o processamento para identificação de perfis e previsão de carga é terceirizado para um ambiente de computação em nuvem. A estratégia proposta modela a transição de usuários entre pontos de acesso como um processo markoviano e, portanto, utiliza cadeias de Markov simples para prever a carga esperada em cada ponto de acesso, de acordo com o perfil de uso dos usuários conectados em cada

momento. Devido ao modelo ser centrado no usuário e em sua mobilidade na rede, analisa-se as transições de associação dos usuários com os pontos de acesso, trazendo, inclusive a possibilidade de previsão sem memória (*memoryless*), uma vez que a única informação necessária é a carga imediatamente anterior na rede.

Por sua vez, o perfil de uso é obtido a partir da análise do tráfego da rede utilizando o algoritmo de aprendizado de máquina para agrupamento não supervisionado *k-means*. A agregação em perfis, além de deixar os algoritmos mais eficiente, contribui para a anonimização dos dados, em conformidade com a Lei Geral de Proteção de Dados (LGPD) brasileira[1] e como Regulamento Geral sobre a Proteção de Dados (GDPR) europeu[2]. A conformação às leis de privacidade do usuário é mandatória à proposta, pois o processamento dos dados ocorre em um ambiente de nuvem pública comercial[3]. Os modelos gerados pela estratégia proposta podem ser usados, também, para detectar anomalias na rede e possibilitar aplicações inteligentes, como cache na borda da rede e garantia de desempenho de serviço personalizada. Além disso, a proposta permite implementar ferramentas de controle que atuem para melhorar a eficiência e a segurança.

A análise da proposta é feita utilizando dados de tráfego obtidos da rede sem fio institucional da Universidade Federal Fluminense (UFF). A rede é composta por 670 pontos de acesso distribuídos em 16 *campi* universitários distintos, com mais de 90 prédios, capaz de atender a mais de 60 mil usuários únicos apresentando picos de acesso de 5.000 usuários simultâneos. Os dados foram coletados por uma semana e representam os fluxos de comunicação entre os clientes na rede sem fio e serviços na Internet referentes a um dos *campi*, comportando 363 pontos de acesso. Os dados coletados foram anonimizados para garantir a privacidade dos usuários. A caracterização do tráfego de dados coletados foi realizada por Magalhães e Mattos [9]. Os resultados obtidos mostram que, na rede analisada, é possível classificar os usuários em 5 perfis de uso e que o mecanismo proposto prevê a carga nos pontos de acesso, considerando um intervalo de confiança de 90%.

O restante deste Capítulo está organizado da seguinte forma. A Seção 5.1 discute os trabalhos relacionados. A Seção 5.2 apresenta o problema de caracterização de perfis de uso da rede e previsão de tráfego nos pontos de acesso. A Seção 5.3 propõe a estratégia de

[1] http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm
[2] https://eugdpr.org/the-regulation/
[3] Os resultados apresentados foram obtidos através da execução na nuvem Amazon Web Services (AWS).

previsão de carga e identificação de perfis. A estratégia é avaliada no cenário de uma rede sem fio de larga escala real e os resultados são apresentados na Seção 5.4. Por fim, a Seção 5.5 conclui o trabalho e aponta trabalhos futuros.

2.1 Trabalhos Relacionados

Trabalhos anteriores focam no estudo da rede através da análise da conexão de novos dispositivos em redes sem-fio [10], na identificação de gargalos de desempenho nas interfaces de rede dos pontos de acesso [7], na caracterização do perfil de uso do tráfego de dados por aplicações [11, 12], na análise do desempenho das aplicações em cenários restritos [13, 14] ou se baseiam em algoritmos de seleção de características para detecção e resolução de incidentes de segurança em tempo real [15].

Boutaba et al. apontam diversas aplicações de aprendizado de máquina em redes de computadores e salientam que existem problemas de natureza discreta, resolvidos por algoritmos de classificação, tais como árvores de decisão, k-vizinhos mais próximos e floresta aleatória, enquanto problemas contínuos são resolvidos por algoritmos de regressão [16]. Nesse contexto, o problema de previsão de carga nos pontos de acesso configura-se como um problema de natureza contínua e, portanto, ao ser modelado como um problema de aprendizado de máquina, deve considerar uma solução baseado em algoritmos de regressão. Contudo, neste trabalho, considera-se a análise do processo estocástico de chegada de novos clientes em detrimento da aplicação de mecanismos de aprendizado de máquina.

A caracterização do perfil de uso em redes Wi-Fi e o correto dimensionamento da rede têm papel fundamental na experiência e na qualidade da rede. Oliveira et al. Investigam e classificam as atividades de usuários em redes universitárias e redes urbanas, concluindo que há uma correlação linear entre o número de sessões e o número de pontos de acesso [13].

Os usuários, no entanto, mantêm-se conectados em poucos pontos de acesso da rede ao longo do dia, como conclui Magalhães e Mattos, cujos resultados mostram que os usuários se associam a poucos pontos durante o dia mesmo com um número substancialmente alto de pontos de acesso próximos [14]. Magalhães e Mattos mostram ainda que a interferência que cada rede gera depende do nível de carga que a rede é submetida, em resultado semelhante ao de Biswals et al. que confirmam que monitorar as

redes sem fio exige monitorar o canal de rádio usado por cada rede Wi-Fi [7], pois o uso do canal pode ser baixo perante as demais redes. Por sua vez, o modelo de coleta de dados adotado neste capítulo segue a abordagem de monitorar os fluxos na rede e consolidá-los com informações de associação de endereço de rede com o ponto de acesso conectado [17].

Ghosh *et al*. realizam a caracterização do perfil de tráfego baseada na chegada de usuários na rede e, portanto, combinam um agrupamento estático com regressão de Poisson para modelar o processo de chegada de novos dispositivos à rede sem fio [10]. Quian *et al*. se baseiam nas aplicações utilizadas pelos usuários da rede para caracterizar e otimizar o tráfego [11]. Os autores propõem uma abordagem entre camadas, em que identificam de tráfego de aplicações populares e otimizam a comunicação de controle de rádio para aplicações que são ineficientes para a camada de rádio, reduzindo gargalos de rede provocados por essa ineficiência. Essas abordagens são sintetizadas por Shye *et al*. Os autores caracterizam o uso da rede sem fio através da observação de que os dispositivos móveis são mais numerosos na rede e que cada aplicação gera uma assinatura distinta de consumo de dados na rede [12]. Diferentemente, a proposta deste capítulo considera a análise dos fluxos de rede através de um agrupamento não supervisionado para identificar perfis de uso da rede e prever a utilização em cada ponto de acesso com base nesses perfis. Abordagem semelhante foi realizada por Lopez *et al*. com foco em detecção de anomalias de segurança, porém com a utilização de algoritmos de classificação supervisionada, que requerem um conjunto de dados conhecidos para treinamento do mecanismo [15].

Para a previsão de carga na rede, deve-se prever também as associações de usuários em cada ponto de acesso. Lyu *et al*. utilizam uma cadeia de Markov de ordem K para prever o próximo ponto de acesso a que o dispositivo se conectará a partir de uma série de conexões anteriores [18]. Diferentemente, neste trabalho consideram-se cadeias de Markov de primeira ordem, pois o objetivo é definir a probabilidade do dispositivo estar associado a um determinado ponto de acesso e não definir o rastro provável realizado pelo usuário. Para validar a premissa de que o ponto de acesso conectado atual só depende do estado anterior e, por consequência, validar o uso de cadeias de Markov simples, considera-se o estado “não conectado" para dispositivos que no período não se conectaram à rede. Abordagem semelhante é utilizada por Mattos et al. na previsão de estado de cada controlador de rede com base na demanda de fluxos de rede [19]. Utilizando-se dos mesmos dados, Rodrigues et al. [20] analisam que a carga de dados na rede da UFF é

bastante concentrada, podendo-se manter apenas 23% dos pontos de acesso ligados enquanto seria capaz de atender 98% dos usuários.

2.2 Problema de Previsão de Carga em Redes Sem Fio de Larga Escala

Em redes sem fio de larga escala, a análise do tráfego de rede para detecção de perfis de uso e previsão de carga é essencial para a segurança da rede sem fio, pois permite assegurar a disponibilidade da rede. Contudo, a previsão acurada da carga é desafiadora devido à variedade de fontes de dados, representadas pelos diversos pontos de acesso e usuários da rede, e devido ao grande volume de dados gerados. O primeiro desafio é a coleta dos dados, realizada tanto nos pontos de acesso distribuídos geograficamente quanto nos roteadores de saída da rede. Os pontos de acesso registram atividades de conexão, atividades de associação e desassociação de clientes, enquanto os roteadores de saída da rede registram resumos do tráfego de rede. Assim, é necessário associar cada fluxo de tráfego de rede com o ponto de acesso utilizado por esse fluxo a cada momento, realizado através da atividade de associação e desassociação dos clientes aos pontos de acesso. Essa associação precisa ser realizada antes de processar os dados coletados. O segundo desafio é a própria análise dos dados, que exige grande disponibilidade de armazenamento e elevado poder de processamento para treinamento dos algoritmos de aprendizado de máquina.

A coleta de dados é feita através da ferramenta *NetFlow*, desenvolvida pela Cisco para monitoramento e padronização de informações sobre a exportação de estatísticas de fluxos de rede pelos roteadores [21]. Os roteadores de saída da rede examinam os pacotes que chegam em suas interfaces e coletam estatísticas dos fluxos com base nas informações da 4-tupla composta por endereços IP de origem e destino, e portas de origem e destino. A criação de um novo registro de fluxo ocorre sempre que se examina um pacote com 4-tupla diferente da do pacote anterior. Todos os pacotes consecutivos que possuem a mesma 4-tupla são considerados como pertencentes ao mesmo fluxo. Os dados coletados são enviados periodicamente para um equipamento de gerenciamento de fluxos, de acordo com temporizadores pré-configurados. O volume de tráfego de rede analisado neste capítulo é da ordem de *gigabytes* por dia e deve ser processado em tempo quase real.

A velocidade e o volume de dados coletados consistem em um desafio para determinar os perfis de uso e a previsão de carga em uma rede sem fio de larga escala,

caracterizando o problema como um problema de *Big Data*. Neste capítulo, propõe-se utilizar uma solução de processamento em nuvem, que permite alocar recursos computacionais sob demanda.

O processamento dos dados é feito conforme eles são coletados, em tempo real. Utiliza-se também a funcionalidade de servidores Spot da nuvem da *Amazon Web Services*[4], que reduz os custos em até 90%, permitindo a criação de uma ferramenta eficiente e de baixo custo [22].

A Figura 6.1 representa a estratégia proposta. Os pontos de acesso e roteadores de borda enviam os dados para um servidor central. O servidor central envia os dados ao ambiente em nuvem *Amazon AWS* utilizando uma rede privada (*Virtual Private Network - VPN*). O ambiente AWS, utiliza um serviço de ingestão de dados para enviar os dados brutos para o *Tier 1*[5], responsável por enriquecer os dados. Os dados enriquecidos são transportados para o *Tier 2* para alimentar o treinamento do algoritmo de agrupamento k-means usado para obter os perfis de usuários. O resultado é salvo no *Tier 3*. Finalmente, os dados enriquecidos e classificados por perfil são usados como entrada para a cadeia de Markov, que aplica o modelo proposto para prever a probabilidade de um usuário de um determinado perfil estar em determinado ponto de acesso.

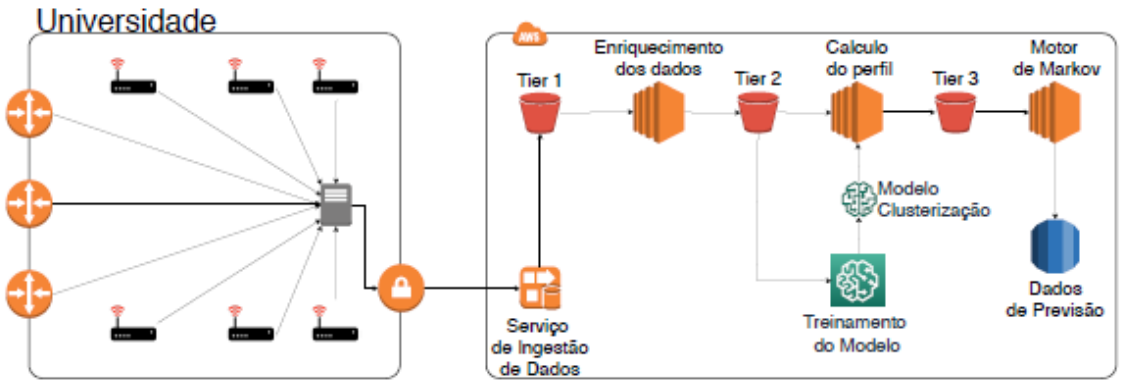


Figura 6.1 - Diagrama da estratégia proposta para previsão de carga na rede e identificação de perfis

A estratégia construída permite uma análise em tempo real, salvo a janela de 10 minutos, caso seja possível coletar os dados em tempo real na rede da Universidade e enviar para a estrutura de processamento na nuvem. No entanto, neste trabalho, devido a

[4] https://aws.amazon.com/ec2/spot/

[5] Tier é a nomenclatura usada em projetos de grande massa de dados para cada etapa de armazenamento de dados em fluxo desacoplado de processamento.

restrições de coleta e envio, foi realizada uma coleta por um período fixo e, posteriormente, enviado para a nuvem para processamento. O modelo é então acionado com os dados a cada janela de 10 minutos, como se fosse uma janela deslizante dos dados, fazendo com que as matrizes sejam atualizadas com base nos dados da janela anterior a cada nova janela de 10 minutos.

2.3 Estratégia para Previsão de Carga na Rede e Identificação de Perfis

Este capítulo propõe uma estratégia para identificação de perfis de uso em redes sem fio e previsão da carga de acesso em cada um dos pontos de acesso [23]. O objetivo é permitir o dimensionamento da rede de forma adequada, de acordo com o perfil de uso de cada localidade e usuário. Para garantir a privacidade dos usuários, os dados coletados são associados a um identificador único aleatório gerado a partir de uma função criptográfica que age sob o endereço MAC do dispositivo. A estratégia é realizada através de implantação de um mecanismo que concilia a captura de dados na rede sem fio e o processamento em uma arquitetura de computação em nuvem. A estratégia consiste em 3 etapas, de acordo com o objetivo final de cada uma. Na primeira, o objetivo é obter a carga esperada em cada ponto de acesso. Na segunda, são identificados os perfis de uso existentes na rede. Por fim, na terceira, identifica-se o comportamento esperado do usuário e o uso esperado em cada ponto de acesso.

Modelo para previsão de carga esperada em cada ponto de acesso

A análise dos dados coletados é feita em intervalos de tempo Δt[6]. Nesse intervalo, avalia-se a associação de cada usuário em cada Ponto de Acesso (*Access Point, AP*) da rede. Um usuário que não realiza nenhuma comunicação de rede no período está no estado "não associado" (AP_0). A probabilidade de um usuário trocar de um AP_i para um AP_j no intervalo de tempo analisado é obtida a partir dos dados coletados, avaliando a proporção entre o número de trocas de associação entre os pontos de acesso e o número total de trocas de associação. A migração de um ponto de acesso para outro é marcada pela presença do fluxo do usuário em um novo ponto de acesso. Considera-se, também, que o usuário migra

[6] Neste trabalho, considera-se Δt = 10 minutos por uma avaliação empírica.

apenas do ponto de acesso atual para o próximo independentemente de qual ponto de acesso estava conectado anteriormente. Logo, o estado atual depende apenas do estado anterior. Assim, a chegada de novos clientes em um ponto de acesso pode ser modelada por um processo de Poisson não-estacionário [10]. Com isso, é possível modelar as transições entre APs utilizando cadeias de Markov. Essas transições preenchem a matriz de probabilidades de Markov.

Neste capítulo, a matriz T_{AP} representa a probabilidade de um usuário trocar de um AP_i para um AP_j no intervalo de tempo analisado. A matriz TAP possui dimensão (N + 1) × (N + 1), em que N é o número de pontos de acesso na rede sem fio, e cada elemento t_{ij}^{AP} representa a probabilidade de transição de conexão de um usuário do AP_i para o AP_j. O estado AP_0 é fundamental para considerar usuários que não aparecem associados a nenhum ponto de acesso no intervalo considerado e atua como um estado vizinho a todos os outros, uma vez que o usuário pode conectar à rede e se desconectar a partir de qualquer ponto de acesso [19]. Assim, para cada usuário da rede, a cada intervalo $\Delta t = m$ de coleta de fluxos, monta-se uma matriz T_{AP}^{m}, como segue:

$$\begin{bmatrix} AP_0 \to AP_0 & AP_0 \to AP_1 & \dots & AP_0 \to AP_N & \\ AP_1 \to AP_0 & AP_1 \to AP_1 & \dots & AP_1 \to AP_N & \\ \dots & \dots & \dots & \dots & \dots \\ AP_N \to AP_0 & AP_N \to AP_1 & \dots & AP_N \to AP_N & \end{bmatrix} \times \begin{bmatrix} AP'_0 \\ AP'_1 \\ \dots \\ AP'_N \end{bmatrix} = \lambda \begin{bmatrix} AP'_0 \\ AP'_1 \\ \dots \\ AP'_N \end{bmatrix},$$

em que $AP_i \rightarrow AP_j$ indica a probabilidade de mudança do ponto de acesso AP_i para o AP_j, e λ é um autovalor correspondente à decomposição da matriz em valores singulares. Escolhendo $\lambda = 1$, tem-se o autovetor $\vec{P}_n^m$ que, normalizado, representa a probabilidade invariante do usuário *n* estar conectado em cada ponto de acesso no instante m_x. Assim, para cada instante m_x, o valor esperado do número de usuários associados ao ponto de acesso AP_k é dado pelo somatório das probabilidades invariantes de cada usuário *n* no momento m_x para AP_k, conforme Equação 6.1.

$$E(Usuários\ em\ AP_k)^m = \sum_{j}^{N} \vec{P}_j^m [AP_k]. \tag{6.1}$$

Dessa forma, pode-se definir a carga do ponto de acesso em função do valor esperado do número de usuários associado a ele ao longo do dia.

Modelo para identificação dos perfis de uso

A estratégia proposta para prever a carga esperada nos pontos de acesso requer elevado poder computacional para processar os dados, mesmo que a análise se baseie somente nos fluxos da rede e nos registros de associação dos dispositivos com os pontos de acesso. Embora a proposta utilize técnicas de *Big Data* para o processamento na nuvem, o volume de dados cresce rapidamente com o número de pontos de acesso e de fluxos na rede, podendo inviabilizar a análise em cenários de menor poder computacional. Assim, com o objetivo de aumentar a eficiência da estratégia, propõe-se realizar o agrupamento dos fluxos em perfis de uso. A estimativa passa a ser realizada em termos de perfis de uso, trazendo uma economia de recursos computacionais, uma vez que os dados de associação ao ponto de acesso podem ser descartados após o processamento e geração dos vetores de probabilidade $\overrightarrow{P}_j^m[AP_k]$ para cada usuário e por cada ponto de acesso da rede.

Para definir o perfil de uso, os dados coletados usando *NetFlow* são enriquecidos pelos dados de associação com o ponto de acesso utilizado em cada instante[7]. Os dados enriquecidos contam com características relacionadas a informações sobre o fluxo na rede, como número de pacotes, duração, quantidades de bytes; a informações sobre o ponto de acesso, como número de usuários associados; e a outras informações categóricas que identificam serviços acessados, protocolos de transporte, *flags* do protocolo de transporte e o ponto de acesso em que cada usuário está conectado. As características categóricas expandem o conjunto de dados para um espaço com mais de 650 características, dado que cada ponto de acesso monitorado transforma-se em uma característica do fluxo. Essas informações são usadas como entrada para o algoritmo de agrupamento não supervisionado *K-means*. O algoritmo tenta encontrar o centroide de grupos discretos *C* dentro dos dados, de forma que os membros de um grupo sejam o mais próximos possível uns dos outros ao mesmo tempo em que se maximiza a distância para os membros de outros grupos. Assim, o objetivo do algoritmo é reduzir a distância interna do grupo e aumentar a distância externa.

[7] O conjunto de dados usado pode ser obtido através de contato com os autores.

Neste capítulo, a distância é o grau de similaridade dos registros, representado pela distância euclidiana. A estratégia proposta utiliza uma versão modificada do algoritmo de agrupamento, *K-means Web-Scale* [24]. Essa versão é mais precisa do que a original, mantendo as características de escalabilidade mesmo para conjuntos de dados volumosos, ao passo que realiza o treinamento em tempo viável. Para tanto, o algoritmo usado utiliza mini-lotes (mini-*batches*) de dados de treinamento, que são conjuntos pequenos e aleatórios dos dados originais. O algoritmo *k-means* recebe os dados em formato tabular, em que as linhas representam os dados de tráfego coletado que será agrupado e as colunas representam as características dos dados. As *n* características em cada linha representam um ponto em um espaço *n*-dimensional.

Por ser um algoritmo não supervisionado, *a priori* os grupos finais do conjunto de treinamento são desconhecidos. No entanto, é necessário passar ao algoritmo o número de grupos que se deseja definir. Cada grupo constitui um perfil de uso da rede e, na análise experimental, apresenta-se um estudo sobre a variação da quantidade de grupos. Com base no número de grupos, durante o treinamento, o algoritmo define um ponto central no espaço *n*-dimensional para cada um dos grupos definidos através da iteração sobre os fluxos de dados, com o objetivo de reduzir as distâncias internas dos grupos e aumentar as distâncias externas dos grupos. A métrica de eficácia para determinar a otimização de parâmetros do modelo e obter a melhor assertividade é definida como a minimização da distância quadrática média de cada ponto em um determinado grupo e o centro do grupo mais próximo. Quanto menor for a métrica de eficácia, melhor é o modelo definido.

Assim, o objetivo é encontrar o conjunto de grupos C dos centroides de grupos $c \in \Re^n$ sendo n o número de características no conjunto de dados e $|C| = k$ o número de grupos ou perfis de uso. O algoritmo minimiza a função

$$min \sum_{x \in X} ||f(C, x) - x||^2, \tag{6.2}$$

através de um conjunto de um conjunto X dos dados coletados, com uma amostra $x \in X$ e $x \in \Re^n$ em que $f(C, x)$ retorna o centro $c \in C$ mais próximo de x usando a distância euclidiana. Como esse é um problema NP-difícil, o algoritmo finaliza após o

número de iterações definido. A quantidade de iterações também é objeto de análise da avaliação experimental da estratégia proposta.

Modelo para previsão de comportamento esperado do usuário e de uso esperado nos pontos de acesso

A terceira parte da proposta prevê a carga em cada ponto de acesso com base no perfil de uso da rede. A partir do agrupamento de cada fluxo e a probabilidade do usuário estar em um determinado ponto de acesso, a previsão considera a probabilidade do usuário ter um fluxo com um perfil de uso definido. Assume-se então que um determinado usuário pode não ter realizado nenhuma comunicação de rede durante o período analisado, estando no estado "sem perfil" (PU_0). De forma semelhante ao modelo para previsão de carga esperada em cada ponto de acesso, é possível montar uma matriz, T_{PU}, em que cada elemento t_{ij}^{PU} representa a probabilidade de um usuário trocar do perfil de uso PUi para o perfil de uso PU_j. Dessa forma, a matriz T_{PU} representa as probabilidades de transição entre perfis de uso e possui dimensão $(K + 1) \times (K + 1)$, em que K é o número de perfis de uso detectados na rede sem fio. Para tanto, assume-se que a transição de um usuário entre diferentes perfis é um processo markoviano e, portanto, independente dos estados anteriores. A hipótese é consistente com a observação de que o perfil de uso da rede está associado às condições de ocupação e de uso instantâneas da rede.

Assim, para cada usuário, monta-se a matriz T_{PU}^{m} de transição de perfil de uso, a cada intervalo de tempo $\Delta t = m$ de fluxos coletados.

De forma semelhante à previsão de carga, considera-se a decomposição da matriz de transição em valores singulares e selecionando o autovetor $\overrightarrow{U}_{n}^{m}$ normalizado, obtém-se o invariante da matriz que representa a probabilidade do usuário n estar gerando um tráfego de rede associado a um determinado perfil de uso no instante m_x.

Como não existe uma correlação histórica entre os fluxos de dados, uma vez que cada fluxo é independente para uma determinada aplicação, é possível utilizar cadeias de Markov simples. Assim, assume-se que o perfil atual depende apenas do perfil anterior, sendo possível a transição de qualquer perfil de uso para qualquer outro perfil. Logo, utilizando as probabilidades de um usuário assumir um perfil de uso e do usuário estar conectado a um dado ponto de acesso, pode-se inferir a probabilidade de um fluxo de um determinado perfil estar utilizando um dado ponto de acesso através da probabilidade

condicionada. Considerando que o usuário estar associado a um determinado ponto de acesso e o seu perfil de uso são variáveis independentes, a probabilidade de um perfil *w* (PU_w) estar presente no ponto de acesso *K* (AP_k) no instante m_x é dada, para cada usuário *j*, com o produto $\overrightarrow{P}_j^m[AP_k] \times \overrightarrow{U}_j^m[PU_w]$. Assim, em cada instante m_x o valor esperado para a carga de cada perfil de uso (PU_w) no ponto de acesso *k* (AP_k) é o somatório das probabilidades dos invariantes de cada usuário *n* naquele instante m_x, conforme Equação 6.3:

$$E(PU_w \ em \ AP_k)^m = \begin{cases} Para \ PU_0 \rightarrow \sum_j^N \overrightarrow{P}_j^m[AP_k] \times \overrightarrow{U}_j^m[PU_0] \\ Para \ PU_1 \rightarrow \sum_j^N \overrightarrow{P}_j^m[AP_k] \times \overrightarrow{U}_j^m[PU_1] \\ \ldots \\ Para \ PU_5 \rightarrow \sum_j^N \overrightarrow{P}_j^m[AP_k] \times \overrightarrow{U}_j^m[PU_5]. \end{cases} \quad (6.3)$$

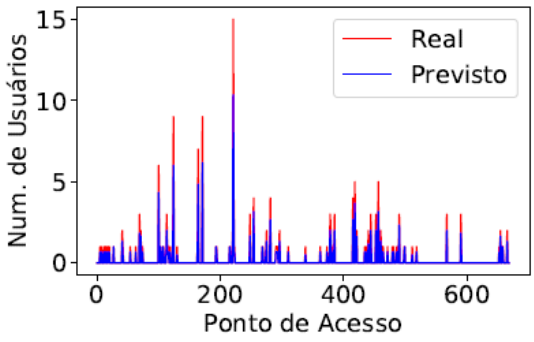
Logo, pode-se definir a carga esperada de cada perfil de uso de rede em cada um dos pontos de acesso ao longo do dia.

2.4 Análise Experimental da Estratégia Proposta

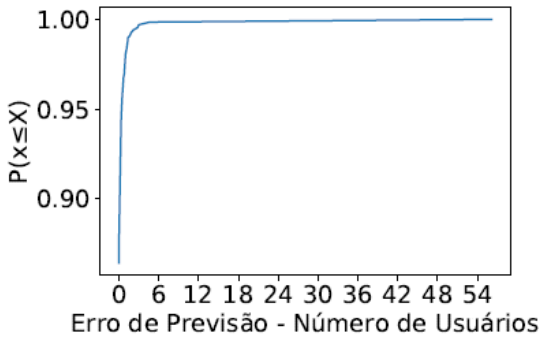
Considerando a estratégia proposta, neste capítulo utiliza-se um servidor virtual na Amazon Web Services (AWS) com 32 GB de memória RAM e 8GB de RAM em GPU para processar todas as matrizes de tráfego da rede acadêmica. A análise da carga em cada um dos pontos de acesso da rede é feita em intervalos fixos. Em uma análise empírica, foi definido uma janela de observação de 10 minutos. Assim, a cada período se avalia a associação de cada usuário em cada AP da rede. A Figura 6.2 mostra que o modelo proposto, baseado em cadeias de Markov simples, para estimar a carga esperada em cada ponto de acesso possui boa assertividade na previsão. As Figuras 6.2(a), 6.2(c) e 6.2(e) apresentam, respectivamente, os resultados obtidos para os dados coletados às 10 h, 13 h e 17 h. Observa-se que cada horário do dia apresenta comportamento de carga diferente, com quantidades de usuários distintas. Além disso, existe grande sobreposição entre os valores para carga prevista e carga real, para todos os pontos de acesso. As Figuras 6.2(b), 6.2(d) e 6.2(f) apresentam a distribuição cumulativa dos erros de previsão do modelo para cada

hora do dia representada pelas Figuras 6.2(a), 6.2(c) e 6.2(e). Observa-se que independente do uso característico de cada hora, o modelo proposto apresenta estimativa dentro da margem de erro de 90% da quantidade de usuários para mais de 90% dos Pontos de Acesso da rede.

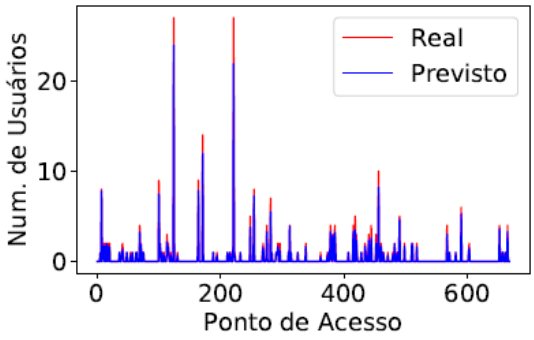
A Figura 6.3 apresenta como o modelo proposto acompanha a curva de quantidade de acesso real para dois pontos de acesso específicos. Os pontos de acesso escolhidos, AP 222 e AP 255 são pontos críticos da rede, pois estão entre os 20 mais utilizados de toda a rede da Universidade. O AP 222 apresenta a pior previsão, maior erro absoluto cumulativo, dentre os pontos de acesso com maior utilização na rede, enquanto o AP 255 apresenta a melhor previsão, menor erro. O comparativo entre a previsão e a carga real é mostrado na Figura 6.3(a) para o AP 222 e na Figura 6.3(b) para o AP 255. Observa-se que o modelo tende a prever os acessos com menor assertividade no início das operações, quando há menor tráfego na rede. Neste cenário, mesmo com o número de usuários baixo, como não há transições na madrugada, o modelo não consegue identificar o erro e mantém os resultados, uma vez que a matriz é atualizada a cada janela de 10 minutos com as transições realizadas. Conforme o número de fluxos na rede aumenta, cresce o volume de dados para análise, melhorando a precisão do modelo para ambos os APs. Essa conclusão é confirmada nos exemplos da Figura 6.2, cuja precisão do modelo é maior para a análise das 17 horas em relação às análises em horários anteriores. Essa observação é consistente para todos os pontos de acesso. Essa característica é favorável ao modelo proposto, uma vez que a precisão da previsão aumenta com o crescimento do número de associações à rede. O modelo proposto é adequado para redes sem fio de larga escala.



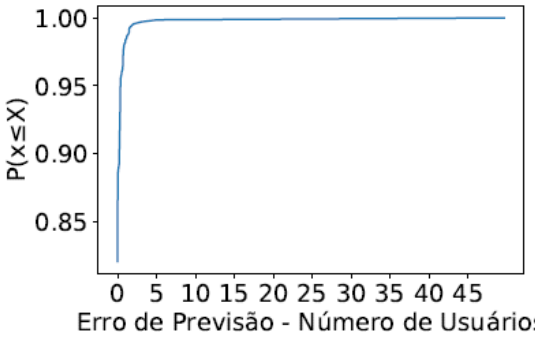
(a) Carga nos APs às 10 horas.



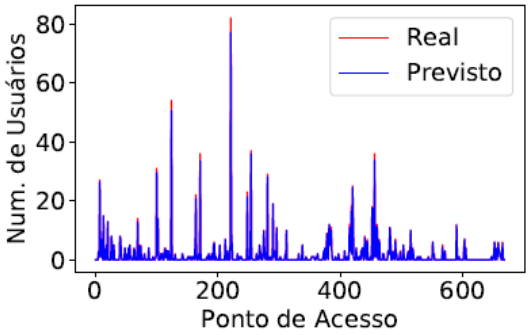
(b) CDF do erro da previsão para as 10 horas.



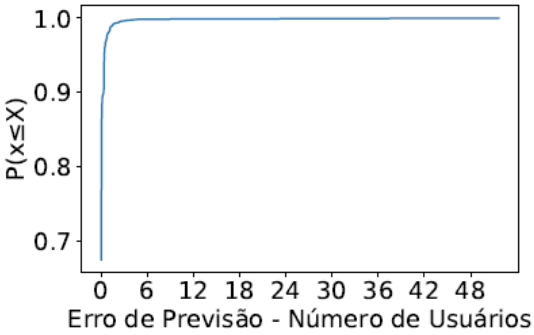
(c) Carga nos APs às 13 horas.



(d) CDF do erro da previsão para as 13 horas.

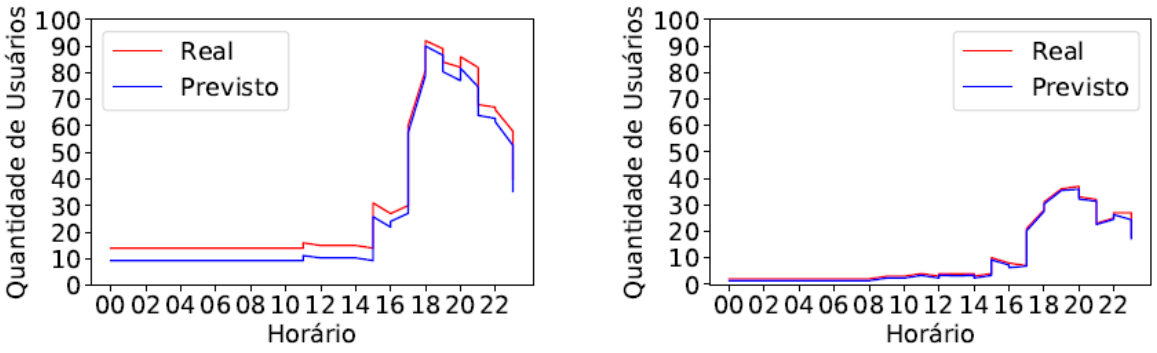


(e) Carga nos APs às 17 horas.



(f) CDF do erro da previsão para as 17 horas.

Figura 6.2 - A comparação entre a carga real e a carga prevista pelo modelo proposto com base nos fluxos de rede e associações com pontos de acesso em três horários distintos, assim como a função de distribuição cumulativa (*Cumulative Distribution Function, CDF*) dos erros de previsão do modelo nos mesmos horários, mostram que o modelo proposto possui boa assertividade na previsão.



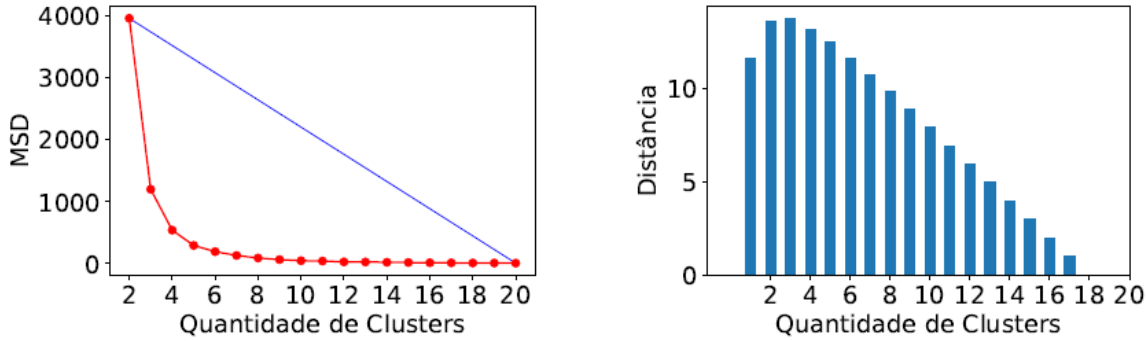
(a) Carga no AP 222. (b) Carga no AP 255

Figura 6.3 - Comparação entre a carga real e a prevista para os Pontos de Acesso 222 e 255 que estão entre os 20 APs mais utilizados. a) O AP 222 apresenta o maior erro absoluto cumulativo da previsão. b) O AP 255 apresenta o menor erro cumulativo absoluto.

A proposta se baseia no algoritmo k-means Web-Scale para identificar os perfis de uso da rede. A Figura 6.4(a) apresenta o comparativo do resultado de erro médio quadrático da distância dos fluxos aos perfis para uma quantidade de agrupamentos que variou de 2 a 20 grupos. Utilizando o método *Elbow*, pode-se concluir que a quantidade de agrupamentos ideal para os dados analisados são 5 grupos. O método *Elbow* visa identificar o melhor balanceamento entre custo de processamento e qualidade do modelo de dados, identificando quando o ganho deixa de ser significativo. Para tal, traça-se uma linha reta de suporte ligando as duas pontas do gráfico de erro por número de grupos, em azul, e calcula-se a distância entre o desempenho do algoritmo com dada quantidade de grupos, curva em vermelho, e essa reta de suporte [25, 26]. A Figura 6.4(b) apresenta que há a inversão na tendência de crescimento do erro, indicando o melhor compromisso entre número de grupos e processamento. Portanto, consideram-se 5 perfis de uso na rede analisada, o que está em consonância com o encontrado por Reis et al [27], que realizou um estudo aprofundado das variações no número de clusters utilizando os mesmos dados coletados. Dessa forma, pode-se aplicar rapidamente, a cada novo fluxo reportado pelo NetFlow, o modelo de agrupamento de fluxos de rede em perfis de uso e, assim, determinar o padrão de carga sendo aplicado pelo usuário no ponto de acesso.

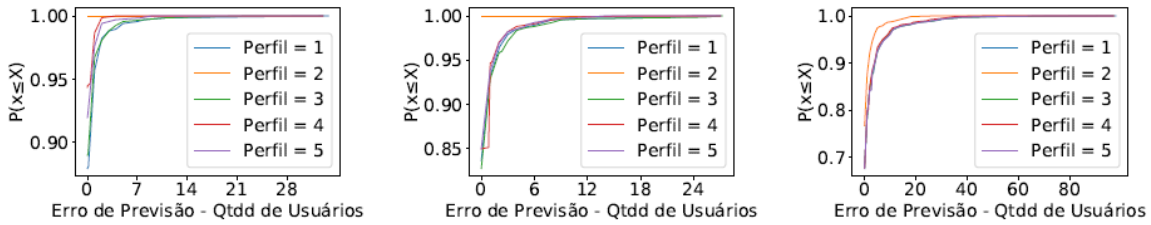
Utilizando o dado enriquecido pelo perfil de uso, pode-se executar a estratégia proposta e prever a carga em cada ponto de acesso por perfil de uso. A Figura 6.5 apresenta as distribuições cumulativas de probabilidade dos erros de previsão para cada um

dos perfis em três momentos do dia: 10h, mostrado na Figura 6.5(a), 13h na Figura 6.5(b) e 17h na Figura 6.5(c). Os resultados revelam que a previsão de carga baseada no perfil é acurada, dado que para todos os perfis o número de erros é inferior 5 para mais de 90% dos casos avaliados.



(a) Método Elbow. (b) Erro quadrático médio.

Figura 6.4 - Representação do número de agrupamentos que melhor representa o conjunto de dados. a) O método Elbow busca o número de agrupamentos, curva vermelha, que mais se afasta da reta de suporte em azul. b) Erro quadrático médio entre as amostras e o centroide em cada configuração com diferente número de agrupamentos.



(a) Carga de perfis às 10h. (b) Carga de perfis às 13h. (c) Carga de perfis às 17h.

Figura 6.5 - Comparação das funções de distribuição cumulativa (CDF) dos erros de previsão da carga do perfil nos pontos da rede às a) 10h, b) 13h e c) 17h.

2.5 Discussão

As redes sem fio de grande escala estão cada vez mais comuns e sua adoção tem crescido consideravelmente. No entanto, prever a carga nessas redes e estimar como os usuários utilizam os diversos pontos de acesso é ainda um desafio em aberto. Esse capítulo propôs a utilização de dados dos fluxos de rede, reportados pela ferramenta *NetFlow*, e dados de associação de dispositivos aos pontos de acesso para a construção de um modelo de previsão de carga nos pontos de acesso por cada um dos perfis de uso.

A estratégia proposta [23] executa com precisão em mais de 90% dos cenários testados na rede sem fio do campus da Universidade Federal Fluminense (UFF), sendo possível prever a carga em cada um dos pontos de acesso da Universidade.

Capítulo 3 - Estudo 2 - Garantia do Privilégio Mínimo em Nuvem Pública

3.1 Computação em Nuvem

A definição formal de computação em nuvem dada pela *Amazon Web Services*, líder do mercado no segmento, é:

> *Entrega sob demanda de recursos de TI e aplicativos, por uma rede pública ou privada, com modelo de definição de preço conforme a utilização.*[8]

Três pontos principais caracterizam o porquê da computação em nuvem de forma diferente do ambiente tradicional [6]. Primeiramente, é importante entender a entrega de recursos de TI sob demanda. Dentre os recursos que são normalmente usados em computação em nuvem estão armazenamento, processamento e memória, por exemplo. É possível contratar a utilização de um ambiente com servidor e banco de dados, utilizar os recursos desejados, aumentar ou diminuir o uso dos recursos, e desligar os recursos conforme necessário. Em relação aos aplicativos, muitos têm contratos mensais de utilização, em vez de um modelo de compra de licenças.

Outro ponto relevante é que os recursos são entregues por uma rede pública ou privada. Para uma solução de computação em nuvem, é muito importante que os recursos sejam acessíveis de qualquer lugar, seja através de uma rede pública ou através de uma rede privada. Uma característica diferenciadora da computação em nuvem é que não são instalados os serviços na infraestrutura local. Os serviços são executados em centros de dados do provedor de nuvem e os clientes acessam aos recursos para integrar a operação.

O modelo de definição de preço conforme utilização difere do modelo de preço de operar serviços em centros de dados. Antes da computação em nuvem, era necessário comprar computadores e equipamentos de rede para construir a infraestrutura de Tecnologia da Informação (TI), com um grande investimento inicial (*Capital Expenditure - CapEx*) e alto custo de manutenção (*Operacional Expenditure - OpEx*), com climatização e energia elétrica. Com a computação em nuvem, não é necessário um investimento inicial

[8] Definição encontrada em: https://aws.amazon.com/what-is-cloud-computing/.

para a construção de infraestruturas complexas e sua manutenção. O custo está diretamente associado à utilização.

3.1.1 Modelos de Negócio

Em computação em nuvem, há três principais modelos de negócios [29], como apresentado na Figura 2.1[9]: (i) SaaS (*Software as a Service*), (ii) PaaS (*Platform as a Service*) e (iii) IaaS (*Infrastructure as a Service*). Esses modelos são a base para entender a responsabilidade do cliente da nuvem ao utilizar os serviços em nuvem, o que está diretamente associado às demandas de segurança que devem ser tratados pelos clientes da nuvem. No entanto, não devem ser entendidos como uma categorização estática, pois diversos serviços apresentam características de mais de um modelo de negócio, tornando a classificação mais próxima de uma definição contínua do que de categorias discretas.

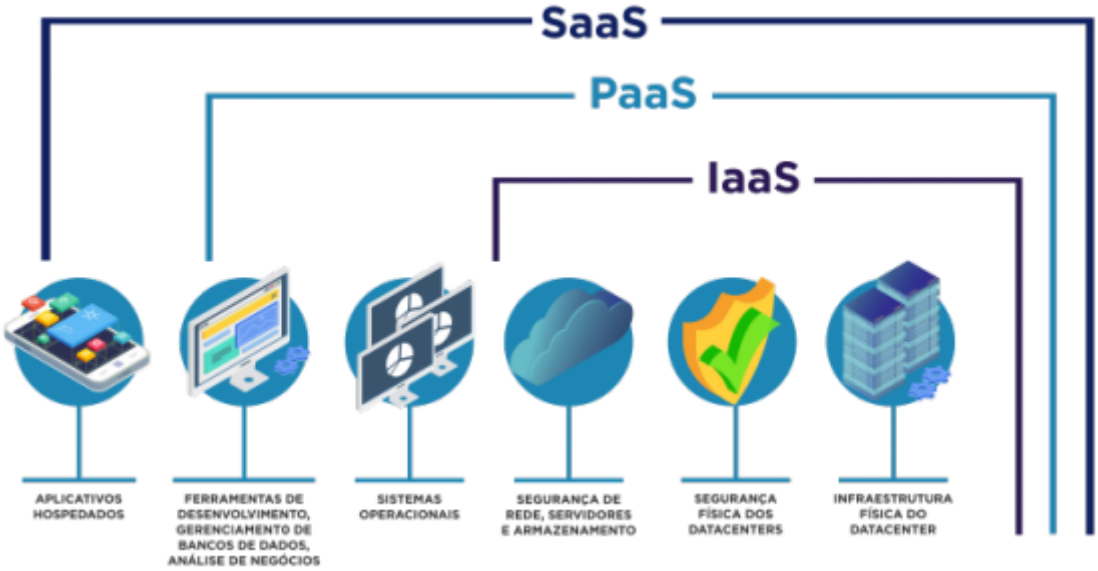


Figura 2.1 - De acordo com o modelo de utilizado na computação em Nuvem (IaaS, PaaS e SaaS), a nuvem provê mais serviços embarcados simplificando as necessidades de configuração e operação.

[9] Imagem extraída de https://solvimm.com/blog/o-que-e-computacao-em-nuvem/.

SaaS (*Software as a Service*)

No modelo SaaS, os serviços na nuvem são software contratados pela Internet, muitas vezes com modelo de contratação mensal, tais como o Dropbox, Netflix e Spotify. As principais vantagens desse modelo para o usuário são a facilidade de acesso aos serviços de qualquer lugar e a facilidade para instalação e atualizações. Já para o fornecedor, os benefícios mais relevantes são o maior controle sobre a sua tecnologia, o combate à pirataria e a consistência na versão utilizada pelos usuários. O contraponto do modelo de assinatura é a forma tradicional de vender licenças de software. Diversas empresas estão migrando para esse modelo como a Adobe[10], com a suíte Creative Cloud, e a Microsoft[11], com o Office 365.

PaaS (*Platform as a Service*)

Já no modelo PaaS, disponibilizam-se plataformas para desenvolvimento, execução e gestão de aplicações. Uma das mais utilizadas é o Heroku, em que o desenvolvedor pode facilmente executar e gerir suas aplicações. Esse modelo de negócios gera muita facilidade e agilidade para o desenvolvimento de novas aplicações e experimentação, fornecendo transparência na gestão de recursos de infraestrutura para os clientes, que não necessitam gerenciar o sistema operacional e a rede dos seus servidores.

IaaS (*Infrastructure as a Service*)

No modelo IaaS é possível contratar armazenamento, processamento, memória e outros recursos de infraestrutura como serviço. Entre os modelos citados, é o que permite maior flexibilidade e gestão de recursos, criação de infraestruturas próprias que atendem exatamente à demanda das aplicações e dos requisitos de compliance das empresas. Em geral, quando as empresas buscam uma rápida redução de custos em TI, uma alternativa simples é migrar datacenters para soluções IaaS.

[10] Extraído de https://fptcloud.com/en/how-adobe-became-a-successful-95-billion-saas-company/.
[11] Extraído de https://www.itassetmanagement.net/2015/06/02/microsofts-move-subscription-saas/.

3.1.2 Mercado de Fornecedores de Nuvem Pública

No mercado, há diversos provedores de IaaS, como a Amazon Web Services (AWS), Microsoft Azure e Google Cloud Platform. A AWS foi a pioneira do segmento e hoje lidera o mercado com ampla vantagem em relação aos seus concorrentes. Gartner, empresa de consultoria norte-americana que realiza pesquisas no mercado de TI, divulga periodicamente um estudo do mercado de computação em nuvem. O quadrante mágico, que identifica os principais competidores do mercado de acordo com a sua capacidade de execução e visão de longo prazo para a tecnologia[12]. A última edição, publicada em Junho de 2022, apresenta a Amazon Web Services (AWS) como líder de mercado e player de maior capacidade de execução dentre todos os avaliados e em empate com a Microsoft na capacidade de visão.

A AWS, por exemplo, oferece mais de 250 serviços, máquinas virtuais Windows, Linux e Mac OS X, bancos de dados relacionais MySQL, PostgreSQL, MariaDB, Oracle e MS SQL Server, bancos de dados OLAP e não-relacionais, serviços de armazenamento e muito mais. Atualmente, é possível contratar recursos computacionais em blocos de 1ms de processamento para códigos em linguagens como Java, Python e NodeJS, entre outras. Os demais provedores, como Azure e Google, têm investido e estão lançando mais recursos em suas plataformas, o que, conjuntamente, apresenta cada vez mais benefícios para as empresas que adotaram esse paradigma de computação em nuvem [30].

3.1.3 Benefícios da Computação em Nuvem

Em relação aos ambientes de centros de dados tradicionais, a computação em nuvem apresenta alguns benefícios que a tornam atrativas para adoção por diversas organizações, comerciais, governamentais e sem fins lucrativos. Nesta seção, destacam-se os principais benefícios.

Preço conforme a utilização

[12] Extraído do relatório encontrado em https://aws.amazon.com/pt/blogs/apn/aws-is-the-longest-running-gartner-cloud-infrastructure-and-platform-services-magic-quadrant-leader/

O pagamento dos serviços de computação em nuvem é cobrado apenas pelos recursos utilizados. Em alguns casos, paga-se por tempo de utilização (segundos ou frações de segundos dependendo do serviço), em outros por espaço de armazenamento, volume de banda de rede ou ainda pelo número de requisições realizadas para a aplicação. Toda a utilização é realizada sem contrato de utilização mínima ou compromisso de uso por tempo determinado, embora contratos desse tipo possam ser realizados para obtenção de descontos.

Dessa forma, as empresas não precisam investir tempo e dinheiro para construir datacenters e manter as estruturas, que inicialmente são mais caras do que o necessário para prever o crescimento das aplicações e do negócio e, posteriormente, se tornam sobrecarregadas, visto que a sua expansão é mais demorada. Portanto, a computação em nuvem é também mais eficiente, gerando apenas o custo necessário a cada momento do crescimento do negócio e das aplicações.

Economia de escala

Os provedores de nuvem possuem estruturas em grande escala em todo o mundo. Isso faz com que sejam grandes compradores dos insumos necessários para a construção e manutenção dos centros de dados, dando a esses provedores grande força compradora. Essa força, promove economia de escala, uma vez que a compra em volume reduz o custo unitário. Como exemplo, a Amazon Web Services (AWS) é a maior compradora de discos rígidos e a maior licenciadora única de Windows no mundo. Isso permite negociações melhores nos diversos aspectos como compra de hardware, contratos de licenciamento de software do sistema operacional e banco de dados e, também, na compra de insumos de segurança, limpeza e até energia elétrica para a infraestrutura computacional. [31]

Elasticidade

Nos ambientes tradicionais, sempre que se compra recursos de TI, deve-se estimar quanto a empresa vai necessitar nos próximos meses e anos, pois o processo de compra e provisionamento dos recursos de TI é demorado. Envolve as etapas de levantamento das opções, escolha da melhor solução e cotação, processo de compra, recebimento dos equipamentos, instalação e configuração dos sistemas para, enfim, integrá-los ao parque

disponível anteriormente. Mesmo em empresas com muita agilidade, esse processo leva semanas.

Com a computação em nuvem, os departamentos de TI não precisam prever a capacidade futura necessária. É possível adquirir em alguns minutos apenas a capacidade necessária naquele momento, aumentando ou reduzindo a oferta de recursos de computação com a demanda por esses recursos exercida pelas aplicações. Dessa forma, os serviços de TI escalam juntamente com o negócio, apoiando a maior eficiência dos recursos. A velocidade com que esse processo pode ser feito, em alguns casos com aumento de milhares de vezes a capacidade em poucos segundos, permite que a empresa possa lidar rapidamente com mudanças bruscas de demanda enquanto atende a todos os seus usuários.

Agilidade para inovar

Esse mesmo processo de alocar e adquirir rapidamente os serviços de TI que a nuvem oferece tem especial benefício na hora de criar produtos e experimentar cenários para inovar no mercado. Dado que é possível provisionar recursos quase instantaneamente e deixar de utilizar os mesmos recursos na mesma velocidade, os departamentos de inovação podem desenvolver e executar experimentos com muito maior velocidade, medindo seus resultados e reagindo também rapidamente a esses resultados. Isso faz com que o ciclo de aprendizado seja acelerado, trazendo agilidade para inovação.

Diversas startups foram criadas e bem-sucedidas ao escolherem a computação em nuvem para impulsionar o seu crescimento, como Dropbox, Uber, Spotify, Netflix. Essas empresas, além de utilizarem dos benefícios de elasticidade e agilidade para inovar da nuvem, puderam focar seus recursos financeiros na construção da sua aplicação e na evolução do negócio, evitando gastos massivos na construção e manutenção de centros de dados, como vistos por startups criadas antes do advento da computação em nuvem.

Com esse foco nas aplicações e na inovação do negócio, essas empresas prosperaram mais, e vemos um grande crescimento no lançamento de novos produtos e em rápido crescimento.

Entrega global

Esse rápido crescimento permite que startups e novos negócios cresçam para atender o mundo todo de maneira muito mais simples e rápida. Assim, há a necessidade de entregar as aplicações com proximidade dos usuários, seja por conta do tempo de carregamento dos serviços, seja por conta de legislação nos países que obriguem os dados dos seus cidadãos a estarem dentro do seu território. Tradicionalmente, seria necessário que a empresa fizesse contratos com centros de dados em diversos países e fosse capaz de gerenciar as diferenças de tecnologia de cada centro de dados além da infraestrutura de telecomunicação que os interligassem [32].

Desacoplamento

Com o aumento da complexidade das aplicações e a velocidade usada pelas empresas para inovação, a agilidade é cada vez mais presente nos times ao redor do mundo, com times independentes que atuam em componentes distintos da aplicação. Essa característica exige que as aplicações sejam segmentadas em componentes desacoplados para que cada time possa trabalhar com independência em sua própria velocidade. Mesmo que os times de desenvolvimento já adotem soluções modulares antes da nuvem, o modelo de infraestrutura centrada no uso de servidores estáticos sempre se tornou um gargalo para a publicação das novas versões dos componentes. No modelo tradicional, exige-se janelas de manutenção para publicação dos componentes nos servidores.

Com a computação em nuvem, a infraestrutura também se tornou desacoplada, uma vez que a interação com os serviços da nuvem também ocorre através de chamadas de interface de programação de aplicações (API). Isso faz com que cada time possa utilizar a sua própria infraestrutura criada através de codificação, fazendo com que cada módulo seja realmente uma aplicação independente. Essa característica da nuvem popularizou o conceito de microsserviços e as aplicações conseguem evoluir com maior velocidade e estabilidade [33].

A estabilidade das aplicações se dá, pois, a falha em um componente não causa necessariamente falha nos demais componentes que interagem com esse. O desacoplamento das aplicações e das aplicações com a infraestrutura, associado ao descarte de servidores com defeito para criação automática de novos servidores saudáveis, possibilita níveis de disponibilidade superiores para as aplicações em nuvem e maior facilidade na investigação e resolução de incidentes. Esse benefício, se explorado e

utilizado adequadamente pelos clientes da nuvem, pode construir aplicações à prova de falhas, entregando alta disponibilidade.

Segurança

Disponibilidade é também um dos pilares da tríade de segurança, juntamente com a confiabilidade e a integridade dos dados e comunicações. Como o provedor de nuvem possui clientes e aplicações de diversos segmentos e níveis de segurança diferentes, ele se torna um alvo para atacantes que desejam explorar falhas na nuvem para obter dados confidenciais ou causar dano a aplicações críticas. Por esse motivo, segurança é prioridade de todos os provedores de nuvem, uma vez que uma falha de segurança da própria nuvem pode criar uma crise de confiança de todo o mercado naquele provedor de nuvem, causando perdas de bilhões de dólares.

O investimento em segurança pelos provedores de nuvem faz com que os clientes da nuvem, independentemente do nível de utilização e valor pago se beneficiem de controles rigorosos e de reconhecimento internacional, simplesmente por utilizarem os serviços da nuvem. Em um provedor de nuvem pública como a AWS, existem mais de 200 certificações e comprovantes de conformidade em regulamentos de todo o mundo, implementados através de mais de 2500 controles automatizados[13] que dificilmente seriam atendidos por infraestruturas de computação tradicional em centros de dados. Esses controles trazem ganhos de segurança para os clientes da nuvem, porém não são exaustivos. Portanto, é importante entender as responsabilidades compartilhadas de segurança no ambiente em nuvem [2]. Além disso, como mostrado em Grant et al [34], as arquiteturas de dados na nuvem devem levar em conta modelos de acesso público não apenas em serviços como também nos planos de dados das aplicações.

3.1.4 Modelo de Responsabilidade Compartilhada

O modelo de Responsabilidade Compartilhada é um dos pontos mais importantes da segurança na nuvem. Esse modelo define a responsabilidade na operação da infraestrutura de cada uma das partes envolvidas: provedor, parceiro, cliente e usuário. Isso se dá devido à atuação na configuração dos diversos componentes para que uma aplicação

[13] Dados extraídos em 20 de junho de 2021 de https://aws.amazon.com/pt/compliance/

funcione adequadamente na nuvem, que é feito de maneira conjunta por essas partes. A Figura 2.3[14] apresenta a divisão entre o que o provedor de nuvem garante e o que deve ser garantido pelos clientes e usuários da nuvem.

Responsabilidade do Provedor

Em um ambiente de computação em nuvem, existem diversas camadas de serviço, interação e obrigações que mantêm um sistema em funcionamento. A camada mais baixa é a infraestrutura física dos centros de processamento de dados que conta com sistemas de tolerância a falhas, alimentação de energia e conexão à Internet com redundância, manutenção contínua de equipamentos e presença em múltiplas regiões geográficas diferentes.

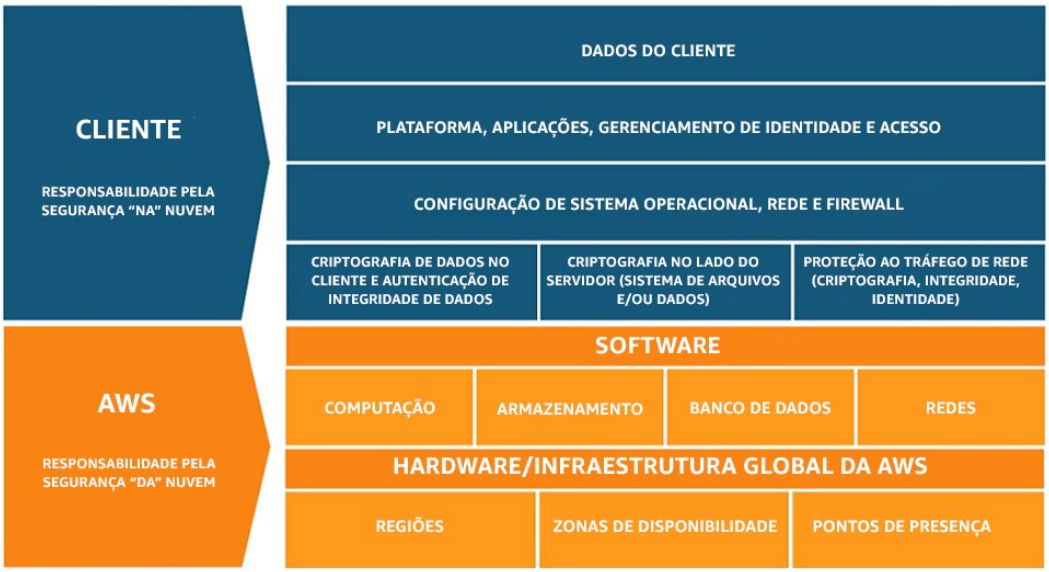


Figura 2.3 - Modelo de Responsabilidade Compartilhada, apresentando as atividades que são responsabilidade do Provedor e do Cliente da nuvem em um cenário de Infraestrutura como Serviço. O Parceiro apoia o Cliente em suas atividades. Fonte: Amazon AWS[14].

Em uma nuvem pública como a Amazon Web Services (AWS), Microsoft Azure e Google Cloud (GCP), logo acima dessa camada, estão os serviços oferecidos e acessíveis pelo painel do usuário, como serviços de computação, armazenamento, banco de dados e rede. Toda a parte de infraestrutura e dos serviços oferecidos é de responsabilidade do provedor de nuvem.

[14] Figura retirada do Blog da Amazon Web Services em https://aws.amazon.com/pt/blogs/aws-brasil/por-onde-comecar-os-estudos-sobre-seguranca-na-aws/

Responsabilidade do Cliente

O Cliente da nuvem é a empresa que contrata os serviços do provedor para implantar suas aplicações. Todo o uso e configuração dos serviços da nuvem para a oferta da aplicação do cliente e seus dados é responsabilidade do cliente. É importante configurar políticas de segurança de senha, por exemplo, que permitam usar apenas senhas fortes com número mínimo de caracteres, números, letras maiúsculas e minúsculas e caracteres especiais. É importante, também, garantir a autenticidade dos usuários, realizando confirmação de criação de conta ou autenticação utilizando uma autoridade confiável, como Facebook, Google ou ainda o diretório centralizado de usuários da empresa [35].

Responsabilidade do Parceiro

O Parceiro é uma empresa de consultoria que oferece conhecimento técnico e mão de obra especializada na configuração e operação dos serviços do provedor de nuvem, permitindo que os Clientes possam focar em construir suas aplicações e desenvolver seu negócio. Para utilizar os serviços da nuvem, é preciso configurá-los conforme as demandas do negócio e, além disso, aplicar medidas de segurança para proteger a aplicação, como criptografia e proteção do tráfego de rede. É necessário configurar o sistema operacional, aplicar atualizações de segurança, políticas de *firewall* e demais configurações. Outra definição muito importante é o controle de acesso a cada uma das partes da aplicação, não só qual usuário pode acessar cada parte do sistema, como também que diferentes serviços utilizados podem comunicar entre si para realizar quais ações.

Essa parte poderia ser responsabilidade do cliente, mas exige um conhecimento muito específico e processos para executar tarefas que desvirtuam do negócio final do cliente. Portanto, para o cliente se dedicar de fato ao que gera valor para o seu negócio, recomendasse que essa atividade seja responsabilidade de um parceiro de consultoria em computação em nuvem, como a e-Core[15].

Responsabilidade do Usuário

[15] https://www.e-core.com

O usuário da aplicação possui responsabilidades centrais ao utilizá-la. Se ele acessar os serviços em uma rede desconhecida e insegura, ou até mesmo em uma rede privada virtual (*Virtual Private Network - VPN*) maliciosa, ele está vulnerável. Além disso, o usuário está sempre sujeito a ataques de engenharia social, em que o atacante busca coletar diversas informações no dia a dia, muitas vezes até se aproximando socialmente da pessoa, para obter os dados que deseja. Por fim, é importante que o usuário utilize sempre senhas seguras, evitando senhas como o seu nome ou de algum familiar.

3.1.5 Desafios de Segurança em Computação em Nuvem

Shahzad [2] defende que o principal desafio das aplicações em nuvem é a confidencialidade dos dados e apresenta os riscos que um cliente da nuvem estar suscetível:

a) **Risco do provedor único de computação em nuvem** - Risco tratado pelo fato de se utilizar um provedor externo que gerencia o ambiente físico do centro de dados da organização, que pode perder o controle dos seus dados e sofrer ataques ou interferência de outras organizações que também são clientes do mesmo provedor de nuvem;
b) ***Economic DDoS*** - Ataques semelhantes aos ataques de negação de serviço distribuídos de aplicações na nuvem, cujo principal objetivo é causar um dano financeiro para a organização atacada, visto que as aplicações escalam para absorver o ataque, o que incorrerá em mais custos;
c) **Segurança do armazenamento em nuvem** - Uma vez que os dados são armazenados em serviços externos de armazenamento, uma má configuração nos aspectos de segurança desses serviços pode levar aos dados serem disponibilizados publicamente.

Shahzad discute as abordagens de segurança da AWS que focam em prover confidencialidade, integridade e disponibilidade para os dados do usuário, é dividida em Certificações e acreditações, Segurança física, Serviços arquitetados para serem seguros e Privacidade dos Dados. O provedor oferece uma série de serviços e funcionalidade para os

usuários garantirem a segurança da sua infraestrutura na AWS, baseado no modelo de responsabilidade compartilhada, e boas práticas de segurança de suas aplicações, que envolvem a proteção de seus dados em trânsito e em armazenamento, a proteção de suas credenciais de acesso, o gerenciamento de múltiplos usuários e a segurança das aplicações. Dessa forma, computação em nuvem ainda possui muitos aspectos para pesquisa no campo da segurança da informação.

Marina et al. [36] identifica os mecanismos de controle de acesso disponíveis nos provedores de nuvem e suas limitações. Shevrin e Margalit [37] apresentam um modelo de detectar ataques a dados armazenados na nuvem através de falhas de configuração de permissionamento. Kumar e Reddy [38] apresentam medidas de segurança fundamentais para garantir proteção de acesso aos dados na nuvem, considerando políticas de controle de acesso, concluindo que a adequação ao princípio do privilégio mínimo é vital para restringir tentativas de acesso não autorizado. Portanto, no cenário de computação em nuvem, o problema de controle de acesso é ainda mais complexo do que em aplicações e em ambientes de datacenter tradicional devido a quantidade de serviços e chamadas de API diversas [39] mapeiam 707 chamadas de APIs diferentes na AWS, que ainda possuem variações de parametrização) e a necessidade de que todos os times acessem o ambiente para realização de rotinas e tarefas, aumentando a complexidade dos permissionamento. Wang et al. [40] mostram os desafios incrementais de aplicações hospedadas na nuvem e as implicações de segurança de serviços expostos externamente, pois os serviços de API são acessados por aplicações mobile e web remotas.

3.2 O Princípio do Privilégio Mínimo na Nuvem

O princípio do privilégio mínimo (*Principle of Least Privilege - PoLP*) consiste em entidades (usuários ou aplicações) privilegiadas atuarem utilizando um conjunto mínimo de permissões necessárias para completar suas tarefas [41]. Os privilégios mínimos protegem contra ameaças como o comprometimento de credenciais, uso indevido acidental, em que entidades com privilégio podem alterar a configuração de algum recurso incorretamente, e uso indevido intencional, no qual a entidade privilegiada abusa de suas permissões para causar ainda mais danos. Desta forma, a implantação do princípio de menor privilégios se mostra desejável e até obrigatório. Sua implementação adequada, no entanto, pode ser difícil e, por vezes, não é aplicada devido à alta carga administrativa. No

que tange ambientes em nuvem, por conta da natureza pública dos endpoints dos serviços e da complexidade dos ambientes, o princípio deve incluir Pessoas, Processos e Tecnologia, conforme definido por Plachkinova e Knapp [42].

Neste Capítulo, são abordados os métodos de controle de acesso aplicados a ambientes em nuvem apresentando uma análise da literatura. Na Seção 3.2.1, apresenta-se o método *Role Based Access Control*, em que as permissões são definidas com base nas ações que os usuários e aplicações executam. A Seção 3.2.2 aborda o método de definição das permissões com base em atributos de usuários, operações, recursos e ambiente. A Seção 3.2.3 apresenta uma especificação dos outros métodos com base em tarefas específicas executadas em uma janela temporal. Por fim, na Seção 3.3 apresentam-se os principais pontos relevantes de informações da literatura para a proposta deste trabalho.

3.2.1 Role Based Access Control

Galante [43] aborda o modelo RBAC (*Role Based Access Control*). Inicialmente, são apresentadas as definições de controle acesso e de dois modelos de controle de acesso seguros e, na sequência, o modelo RBAC é exposto, sendo feita sua avaliação. Em seguida, alguns exemplos que mostram o contraste entre o modelo de controle de acesso simples e o RBAC são abordados, permitindo esclarecer os cenários mais apropriados para a implementação do RBAC. O modelo de controle de acesso simples significa a atribuição de permissões para as ações diretamente para os usuários. Também é abordada a transição do modelo simples para o RBAC, mostrando alguns passos básicos para esta implementação. Por fim, uma discussão sobre a criação de grupos de recurso é feita, mostrando sua relevância ao se aplicar o RBAC.

Galante apresenta a definição de controle de acesso como sendo a modelagem de um conjunto de regras de um sistema de computador seguro. Os três principais elementos envolvidos são: Autoridade, Atributos e Regras. Autoridade se refere ao agente que deve determinar a política de segurança, identificar informações de segurança que sejam relevantes e conceder valores a determinados recursos controlado. Os Atributos especificam as características e propriedades dos sujeitos (pessoas ou entidades) e objetos (recursos ou dados), sendo que o sistema toma decisões a respeito do controle de acesso com base nesses atributos. Já as Regras, consistem em expressões formais que estabelecem as relações entre os atributos e outras informações de segurança para as decisões referentes

ao controle de acesso, transmitindo as políticas definidas pela Autoridade. O controle de acesso se refere à capacidade de conceder ou negar a utilização de um determinado recurso por parte de uma entidade particular. É proposta ainda a substituição do termo entidade por pessoa e da regra de permissão por exceção, o que conduz à redefinição de controle de acesso como o ato de restringir o direito de uso de um determinado recurso, sendo a permissão concedida por exceção. Desse modo, tem-se que o RBAC implanta a segurança com base em regras de negar todos e permitir por exceção.

Justifica-se a implantação do controle de acesso através de três princípios básicos de segurança da informação, que são Separação de Deveres (*Separation of Duties - SoD*), Princípio do Privilégio Mínimo (*Principle of Least Privilege - PoLP*) e Necessidade de Conhecimento (*Need to Know - NtK*).

Separação de Deveres - refere-se ao princípio de que nenhum usuário deve receber privilégios suficientes para usar indevidamente o sistema por conta própria.

Princípio do Privilégio Mínimo - refere-se a um conceito de segurança da informação no qual um usuário recebe os níveis mínimos de acesso – ou permissões – necessários para desempenhar suas funções de trabalho

Necessidade de Conhecimento – refere-se à garantia de que a informação esteja disponível e entregue apenas para pessoas que precisam saber dessa informação.

O PoLP traz uma ideia similar ao NtK, que seria a de conceder os privilégios necessários à realização do trabalho. Também são apresentados os conceitos de controle de acesso discricionário (*Discritionary Access Control - DAC*) e de controle de acesso obrigatório (*Mandatory Access Control - MAC*), os quais são submodelos que implantam o SoD e o PoLP. O DAC é aplicável ao acesso confidencial ou de nível de acesso único, pois concede às entidades o controle de propriedade dos privilégios de acesso de modo que, os proprietários podem conceder objetos a terceiros. O MAC provê o acesso classificado a sujeitos e objetos, com agrupamento utilizando rótulos. Neste modelo, a leitura ou escrita de um certo objeto será permitida ou bloqueada de acordo com a permissão atribuída ao seu agrupamento.

Em contraponto, existe o Controle de Acesso Simples (*Simple Access Control - SAC*), que possui duas formas. A primeira trata-se de uma lista de objetos acessíveis associados a um sujeito (forma normal), a qual é referida também como modelo baseado em capacidade, cuja relação de controle de acesso é: sujeito [pessoa] → objeto [recurso]. A segunda consiste em uma lista de sujeitos privilegiados associados a um objeto (forma

invertida), referida como o modelo baseado em ACL (*Access Control List*), em que a relação de controle de acesso estabelecida é: objeto [recurso] → sujeito [pessoa]. As ACLs identificam assessores legais, isto é, pessoas com permissão para acessar um objeto e, na maior parte das vezes, estão atreladas ao objeto de interesse. Em ambas as formas, normal e inversa, os relacionamentos são de um para muitos, pois a instância primária é isolada relacionalmente de outras de mesma classe.

O SAC pode ser aplicado com eficiência em organizações pequenas, pois a abordagem direta funciona bem para permissões únicas em que não se baseiam em funções. No entanto, à medida que as organizações crescem, a implantação do SAC pode falhar, deteriorando o controle de acesso que antes era eficiente. O RBAC atenua o problema pois, assim como o SAC, restringe o acesso, porém consegue se adaptar melhor às organizações complexas e em expansão. Isto ocorre pois o RBAC não implementa uma relação direta de pessoa-recurso, mas conecta uma pessoa à uma função, que, por sua vez, possui conexão com o recurso. Desse modo, o modelo padrão do RBAC é: *pessoa* ↔ *função* ↔ *recurso*. De forma a simplificar as associações, é proposta a utilização de uma nova classe chamada tipo de recurso, que atua como substituta para um grupo de recursos ativos. Utilizando este conceito, produz-se o conjunto de relacionamento *pessoa* ↔ *função* ↔ *tipo de recurso* ↔ *recurso*. Isto ajuda a reduzir a carga de trabalho, pois contribui, por exemplo, com a configuração mais rápida de objetos que apresentam os mesmos recursos funcionais.

No modelo RBAC, tem-se um fluxo bidirecional, que contrasta com o fluxo unidirecional do SAC, mas que atende aos modelos normal e inverso. No RBAC, as funções possuem o papel de substituto ou de mediadores entre as pessoas que as possuem e os recursos de que estas pessoas necessitam. Os relacionamentos muitos para muitos, que atua no RBAC, atendem ao mundo real pois, geralmente, tem-se pessoas exercitando mais de uma função e funções que compreendem à muitas pessoas. Isto também se estende ao relacionamento entre funções e recursos. Apesar disso, o RBAC também é capaz de atender restrições de pessoas designadas, funções e recursos a um relacionamento um-para-um com objetos vizinhos. Desse modo, o modelo de controle de acesso controla recursos confidenciais ou valiosos, aqueles que têm permissão para acessá-los e as funções que conectam e, além disso, comporta associações fora da banda, que são exceções ao modelo fundamentado em funções.

O RBAC pode ser vantajoso ao ser empregado em pequenas organizações quando há necessidade de controle estrito sobre a acessibilidade devido à sensibilidade dos recursos, classificação ou natureza proprietária. Benefícios provenientes da implantação de RBAC, e que servem de justificativa para a sua utilização, são agilidade, economia e segurança. RBAC proporciona a redução nos custos de algumas formas. A primeira delas é a diminuição da carga de trabalho gerencial empregada na concessão de autorizações por parte dos executivos. Com o uso de funções mediadoras entre pessoas e recursos, uma autorização que precisaria ser concedida à 10 usuários, para 12 recursos diferentes, gerariam 120 aprovações a serem feitas no método SAC. Porém, com o uso do RBAC, demandaria apenas 22 autorizações. A segunda forma de reduzir custos é a redução de erros inerentes de um espaço de tarefa significativamente simplificado e menor. Uma outra maneira de reduzir custos é através da recuperação de custos em recursos desnecessários. Sem o RBAC, estes recursos só poderiam ser desprovisionados quando um funcionário terminasse. Por fim, o sistema RBAC com interface e automatizado proporciona economia financeira e tempo ao substituir as atualizações manuais de ACLs por um processo automatizado, eficiente e limpo.

Um segundo benefício do RBAC é a redução de risco, a qual é aplicada por meio de três ações de controle de segurança: preventiva, detectiva e corretiva. O RBAC evita violações ao limitar o acesso a recursos para pessoas com necessidade verificada. As trilhas de auditoria providas pelo RBAC ajudam na detecção de violações de segurança, proporcionando a extinção da viabilidade contínua de privilégios de acesso. As condições que mudam rapidamente podem gerar vulnerabilidades, então o RBAC pode assinalar estas condições para ação corretiva.

O terceiro benefício trata-se do aumento da responsabilidade e do controle. Para conceder um controle de acesso adequado, o RBAC deve ser capaz de produzir funções, recursos e relatórios de acesso de forma prática, automática e sob demanda, direcionados à pessoas com a incumbência de supervisão. Por ser responsável por pessoas, funções, recursos e relacionamentos entre eles, ele pode determinar um controle de inventário em tempo real (*Just-in-Time*) sobre as permissões. Assim, as pessoas recebem o acesso de que necessitam para executar seus trabalhos no tempo certo, nem antes do necessário e nem por um mais tempo. O acesso pode ser concedido de forma automática à medida que uma pessoa é associada à uma função e revogado automaticamente quando for desassociada

desta função. Desse modo, o RBAC remove as permissões obsoletas ao mesmo tempo que provê os recursos necessários às pessoas de forma rápida e eficiente.

Li et al. [44] tratam os riscos de segurança na computação em nuvem, mantendo o foco no controle de acesso. Os autores apresentam um mecanismo de proteção para a nuvem baseado no RBAC. Os autores utilizam notações de conjuntos e mostram a dinâmica envolvida através de um diagrama UML. A partir dele, são especificados os relacionamentos entre usuários, funções, sessões, administradores e objetos protegidos. É apresentada a diferença entre o RBAC e o DAC (*Controle de Acesso Discricionário*). No RBAC, os usuários não podem passar suas credenciais deliberadamente a outros usuários, diferentemente do DAC. Assim, o RBAC pode ser visto como uma forma de MAC (*Controle de Acesso Obrigatório*) sem requisitos de segurança multinível, em que as regras de permissões são aplicadas diretamente nos sujeitos ou objetos, sem agrupamento de sujeitos ou de objetos. O modelo RBAC simplifica o tratamento do permissionamento em nuvem, pois descarta a possibilidade dos usuários criarem outros usuários com permissões mais privilegiadas.

É possível empregar o modelo RBAC em ambientes da nuvem com sucesso, mas para isso é preciso identificar com clareza as respectivas entidades. Os autores descrevem cada entidade, começando por usuários/agentes. Esta entidade se refere a pessoas ou componentes de software, hardware ou rede dentro da nuvem, podendo ser classificados de acordo com suas funções de trabalho. Já as funções (*Roles*) são classificadas de acordo com as incumbências de trabalho. As funções podem ser divididas de diversas maneiras, sendo a primeira delas de acordo com os acessos característicos, como o acesso a programas, dados e servidores. A segunda maneira é a partir dos três modelos de serviço (SaaS, PaaS e IaaS). Na terceira maneira, segue-se a arquitetura de nuvem determinada pelos provedores, na qual as funções podem ser separadas em níveis mais refinados de modo que, a estrutura de funções forma uma hierarquia. Nessa hierarquia, as funções têm a possibilidade de herdar permissões e funções definidas em funções pai. As permissões são definidas de acordo com as funções de trabalhos exercidas pelas funções. São comumente classificadas em três grupos: permissões de acesso a dados, permissões de acesso a programas e permissões de acesso a serviços. As permissões de acesso devem ter suas configurações implementadas de acordo com os requisitos ambientais. Já os objetos protegidos se referem a recursos dentro da nuvem, e se dividem em três grupos: dados, programas e serviços.

A implantação do RBAC é específica do fornecedor de nuvem, devido à falta de padrões e práticas recomendadas, sendo necessário um investimento por parte dos grandes consumidores para que haja um mapeamento entre as funções de usuários e as funções de negócios internas. Os autores afirmam a necessidade de se ter uma abordagem que permita que funções corporativas possam ser separadas das funções definidas pelos provedores de nuvem. O modelo RBAC pode não ser aplicável em todos os domínios de segurança da computação em nuvem, sendo mais adequado para situações em que as funções de trabalho de diferentes roles podem ser determinadas com clareza e separadas, e as funções possuam uma característica orientada a objetos, podendo formar uma hierarquia.

Para reduzir a carga administrativa, pode-se utilizar métodos de mineração de funções, que criam políticas RBAC a partir de privilégios já existentes. Nesse caso, o *Role Mapping Problem* (RMP) utiliza um conjunto mínimo de funções como a medida de eficácia para os algoritmos de mineração. Sanders e Yue [45] abordam os desafios encontrados na obtenção de privilégio mínimo no ambiente de nuvem, bem como os benefícios potenciais da utilização de métodos automatizados para auxiliar a criação de políticas de privilégio mínimo a partir de logs de auditoria. O trabalho aborda um problema diferente do RMP, pois tem o objetivo de reduzir a carga do administrador criando políticas seguras e completas, e não políticas de fácil gerenciamento. Foram implementadas duas estruturas: a primeira consiste em uma Estrutura de Geração de Políticas que, a partir de logs de auditoria coletados previamente, cria políticas automaticamente; a segunda trata-se de uma Estrutura de Avaliação, que se destina a quantificar a segurança fornecida pelas funções geradas pela primeira estrutura. Os mecanismos citados foram aplicados a um conjunto de dados de log de auditoria do mundo real, com 4,3 milhões de eventos. Os resultados mostram a eficácia do método apresentado em reduzir o excesso de privilégios e a carga administrativa envolvida no gerenciamento de permissões.

Com o intuito de demonstrar os desafios envolvidos na criação de políticas de privilégio mínimo, inicialmente, foram analisados dados de logs do *AWS CloudTrail* com 4,3 milhões de eventos durante um período de 307 dias. Os dados foram comparados aos registros do *AWS Identity and Access Management (IAM)* referente a conta AWS em análise de acordo com a existência ao final do período. Foram abordados dois tipos de entidades: usuários e instâncias de máquina virtual. Nesse cenário, os usuários receberam acesso irrestrito e as máquinas receberam funções criadas manualmente por

administradores da rede. Os resultados mostraram que, mesmo possuindo conhecimento sobre o funcionamento da aplicação cujos logs foram utilizados e das permissões necessárias, os administradores acabaram por criar políticas com excesso de privilégio, nas quais havia uma diferença significativa entre o número de ações concedidas e o número de ações de fato usadas.

Para gerar as políticas, o processo se inicia no consumo de logs de auditoria de dados brutos. Em seguida, os logs consumidos são normalizados criando-se uma projeção dos eventos em cada entidade privilegia identificada (usuário ou instâncias). Então, o algoritmo gerador de política é aplicado aos logs de auditoria normalizados. O algoritmo faz uso da contagem simples, criando concessões de política para cada ação bem-sucedida que uma entidade realizou durante o período de observação. A estrutura de avaliação gerada simula a ação de um gerador de políticas de privilégio mínimo automatizado em diversos períodos de observação e de operação. A estrutura utiliza a abordagem da janela deslizante, gerando repetidamente fases de observação e operação com tamanhos predeterminados e comparando as políticas geradas na fase observação com os privilégios exercidos durante a operação. Os testes ajudam a determinar quanto tempo deve durar a fase de observação.

Para avaliar a eficácia dos mecanismos propostos considerou-se os requisitos fundamentais do PoLP, que são a minimização do excesso de privilégios e a minimização dos sub-privilégios. Para tal, utilizou-se as métricas Precisão, *Recall* e *F-score*, a partir dos quais foi possível identificar a importância de se estabelecer um período de observação adequado. Além disso, abordar entidades diferentes de forma separada se mostrou necessário, pois estas apresentam comportamentos distintos quanto à previsibilidade.

Em um trabalho seguinte, Sanders e Yue [46] apresentam a definição formal para o Problema de Minimização de Erro de Privilégio (PEMP), que se refere ao balanceamento entre privilégio excessivo e privilégio insuficiente e, além disso, apresenta um método que pontua quantitativamente as políticas de segurança. Os autores implementam e comparam três tipos de algoritmos para geração automática de políticas, sendo um algoritmo ingênuo, um algoritmo de aprendizado não supervisionado e um algoritmo de aprendizado supervisionado. Os resultados apresentados se referem a avaliação obtida ao aplicar os algoritmos citados em um conjunto dados do mundo real com 5,2 milhões de entradas de log de auditoria da AWS.

Para tratar o PEMP, que é um problema de previsão, os autores utilizam um algoritmo não supervisionado e outro supervisionado para minerar os dados de logs de auditoria de serviços em nuvem, possibilitando abordar também um algoritmo ingênuo para comparação. Como métrica de avaliação foi utilizado o *F-Measure*, que é comumente empregado para pontuar problemas de classificação binária. Essa métrica, através do parâmetro, permite obter uma pontuação ponderada entre privilégio excessivo e privilégio insuficiente. Desta forma, os resultados apresentados neste trabalho utilizaram um alcance de valores de β com o intuito de demostrar como uma organização pode selecionar a melhor abordagem para a sua realidade a partir de seu nível de risco aceitável.

Foram utilizados dois métodos de aprendizado de máquina para gerar políticas de privilégio a partir de dados de auditoria de mineração. O primeiro deles, foi utilizado para realizar um agrupamento com o intuito de encontrar entidades privilegiadas que usassem permissões semelhantes, similar ao problema de encontrar documentos semelhantes em um texto. Em seguida, gerou-se políticas que combinavam os privilégios utilizados pelas entidades agrupadas. Posteriormente, utilizou-se o segundo método de aprendizado, que é a classificação. A partir do conjunto de relações usuário-privilégio obtido durante período de observação (*Observation Period - OBP*), através da aplicação do primeiro algoritmo, o classificador foi treinado para aprender quais relações usuário-privilégio deveriam ser classificadas como concessão e quais deveriam ser classificadas como negadas. Após a fase de treinamento, o classificador foi empregado na geração de políticas para um período de operação (*Operation Period - OPP*).

Dois tipos de previsões foram abordados: previsões individuais e previsões múltiplas. No caso das previsões individuais, apenas um período de operação (OPP) é considerado, utilizando, portanto, somente um conjunto de teste. Para as múltiplas previsões, os resultados preditos também foram comparados com os alvos, no entanto, o conjunto de dados foi separado em vários conjuntos de treinamento e de teste, considerando mais de um período de operação. Para isto, utilizou-se amostragem para separar os conjuntos, empregando uma abordagem que leva em consideração uma dimensão de tempo com interdependência entre os dados. Esta abordagem é denominada "amostragem fora do tempo", em que se seleciona dados de um período para compor o conjunto de treinamento, e dados de um outro período para compor o conjunto de teste.

O primeiro algoritmo abordado no trabalho se destinava a geração de política ingênua. Para este tipo de política, o algoritmo coleta todos os privilégios exercidos

durante o período de observação e os combina para gerar uma política que será utilizada no período de operação. O segundo algoritmo foi empregado para a geração de políticas não supervisionadas. Neste caso, foi utilizado um algoritmo de agrupamento para que fossem encontrados clusters com entidades privilegiadas similares a partir das permissões que elas exercem. Cada permissão exercida por uma entidade foi adicionada em um documento a parte e, posteriormente, estes documentos foram agrupados. Após agrupar as entidades semelhantes, cada grupo recebe uma função compartilhada com permissões combinadas de todas as respectivas entidades. As entidades que não possuírem um cluster receberão somente os privilégios que elas exerceram durante o período de observação. O terceiro algoritmo foi projetado para criar políticas supervisionadas. Inicialmente, um conjunto de documentos foi construído a partir das permissões exercidas na fase de observação e um subconjunto de dados foi selecionado para criar os rótulos das classes. Na sequência, o classificador foi treinado utilizando o conjunto de treinamento para cada permutação dos Parâmetros do Algoritmo Classificador (CAP). As múltiplas instâncias com diferentes permutações são utilizadas para a seleção de hiperparâmetros. Em seguida, criou-se um conjunto de permissões possíveis durante o período de operação com base nos Parâmetros de Geração de Política (PGP). Cada permissão de política é aplicada no classificador, que faz a previsão determinando se a permissão deve ser concedida ou negada. Os resultados desta classificação são utilizados para gerar outras políticas para o próximo período de operação. Nessa fase, utiliza-se um algoritmo de classificação de árvore de decisão (*Decision Tree - DT*).

Os autores apresentam resultados de acordo com as abordagens feitas ao se estimar os resultados obtidos. De forma geral, os autores mostram que o algoritmo supervisionado apresenta um bom desempenho para reduzir o excesso de privilégios, enquanto os algoritmos não supervisionados mostraram ter um bom desempenho na redução de privilégios insuficientes, comparado ao algoritmo ingênuo. Esses resultados demonstram os benefícios e o potencial que a aplicação dos métodos abordados pode fornecer à criação automática de funções seguras baseadas no nível aceitável de risco que uma organização dispõe. A geração automatizada de políticas sugerida utiliza recursos como nome do serviço, nome do usuário e privilégio exercido, obtidos de dados de logs de auditorias.

3.2.2 Attribute Based Access Control

Sanders e Yue [47] adotam a abordagem de mineração de regras para gerar automaticamente políticas ABAC (*Attribute Based Access Control*) responsáveis por minimizar o privilégio excessivo e o privilégio insuficiente. Segundo os autores, o ABAC tem se tornado popular devido a granularidade, flexibilidade e usabilidade, permitindo que sejam criadas políticas que se baseiam em atributos de usuários, operações, recursos e ambiente.

O trabalho propõe um algoritmo de mineração de regras, aplicado em registros de auditoria para gerar políticas ABAC que minimizam o subprivilégio e o superprivilégio, um algoritmo de uma pontuação para avaliar as políticas criadas e métodos para otimizar o desempenho, capazes de lidar com extensos espaços de privilégios ABAC. Para avaliar a solução proposta, os autores utilizaram um conjunto de dados do mundo real com 4,7 milhões de eventos de registro de auditoria, obtidos a partir do AWS *CloudTra*il.

As políticas de controle de acesso visam determinar quais entidades privilegiadas podem exercer funções sobre objetos a partir de condições. O excesso de privilégio pode provocar danos mais elevados nos sistemas devido a credenciais comprometidas, ameaças internas e uso indevido acidental. Já o privilégio insuficiente impede que usuários e serviços realizem suas funções. Sanders e Yue apresentam o Princípio do Privilégio Mínimo (PoLP), que é um princípio de controle de acesso fundamental na segurança da informação, demanda que toda entidade privilegiada opere em um sistema com um conjunto mínimo de privilégios que possibilitem a execução completa de seu trabalho.

O trabalho traz a definição do problema de minimização de erro de privilégio no modelo ABAC ($PEMP_{ABAC}$). A partir de um conjunto de todas as combinações de *atributo:valor* válidas, determine o conjunto de restrições de *atributo*:*valor* que minimiza os erros de privilégio excessivo e privilégio insuficiente para um dado período de operação (*Operation Period - OPP*). Essa definição exclusiva do $PEMP_{ABAC}$ se faz necessária pois, para criar as políticas ABAC, são considerados como atributos de usuários, operações, recursos e ambiente, o que proporciona um espaço de privilégio mais amplo do que o considerado para o modelo RBAC, de onde é considerada a definição original do PEMP [46].

O trabalho aborda duas modalidades para avaliar o algoritmo proposto. A primeira delas considera a pontuação das previsões individuais. Quando se aplica políticas concebidas a partir de dados do período de observação aos privilégios exercidos no período

de operação, é possível classificar as previsões em uma dentre quatro categorias, que são: Verdadeiro Positivo (TP), Verdadeiro Negativo (TN), Falso Positivo (FP) e Falso Negativo (FN). A partir das respostas anteriores, calcula-se a taxa de verdadeiro positivo (TP_{rate}), conhecida como Recall, e a taxa de falso positivo (FP_{rate}), em que TP_{rate} se representa subprivilégio e FP_{rate} representa privilégio excessivo.

A segunda modalidade de pontuação considera políticas de pontuação em vários períodos de tempo. Neste caso, utiliza-se a validação fora amostra, na qual uma parte dos dados é utilizada como conjunto de treinamento, referido como Período de Observação (OBP), e a outra como conjunto de teste, referido como Período de Operação (OPP). Nessa modalidade, as interdependências temporais entre as ações são conservadas. A Pontuação de identificadores de recursos possíveis infinitos, na qual a quantização de recursos disponíveis é feita através da contagem de identificadores de recursos existentes no OBP e no OPP.

O algoritmo de mineração proposto no trabalho funciona de forma iterativa. Os logs de acesso descobertos são utilizados para criar iterativamente regras candidatas. Para gerar regras candidatas, utiliza-se o algoritmo FP-*growth* [48], que obtém conjuntos de itens frequentes no conjunto de entradas de registro do período de observação descobertas.

Os itens destes conjuntos trata-se de declarações do tipo *atributo*: *valor*. As regras candidatas geradas são pontuadas utilizando a métrica C_{score} apresentada pelos autores. Após pontuar cada regra candidata, a regra com pontuação mais elevada é selecionada e adicionada à política e, na sequência, os itens de registro cobertos por esta regra são removidos do conjunto de entradas de registro não descobertas. Este processo se repete até que todas as entradas de log sejam atendidas.

3.2.3 Task Role Based Access Control (T-RBAC)

O modelo de controle adotado pelo RBAC e pelo ABAC é um modelo passivo, em que os direitos de acesso estão atrelados às funções/atributos e os usuários são atribuídos às funções/atributos apropriados. Esse modelo não é eficiente ao capturar responsabilidades de usuários para tolerar fluxos de trabalho em que a ativação de acessos para determinadas tarefas é dinâmica. O modelo de controle de autorização baseado em tarefas (*Task Based Access Control - TBAC*) é um modelo ativo sem separação entre funções e tarefa. As permissões são ativadas e desativadas a partir da tarefa atual ou do estado do processo. No

Controle Baseado em Tarefas de Funções (*Task-Role-Based Access Control - T-RBAC*), que é um modelo de controle de acesso ativo e que fornece herança parcial de autoridade. Os usuários se relacionam com as permissões através das funções e das tarefas.

Narayanan [49] propõem uma adaptação do T-RBAC para os sistemas de saúde provisionados na nuvem, abordando os requisitos de controle de acesso voltados para sistemas em nuvem multilocatários de saúde e sugerindo uma adaptação do modelo de controle de acesso baseado em funções como privilégio mínimo, separação de tarefas, delegação de tarefas e acesso espacial e temporal. Boomija e Raja [50] estendem o conceito de aplicação na área da saúde utilizando criptografia homomórfica para encriptar os dados e adicionar mais uma camada de proteção.

O controle de acesso aos dados deve ser flexível e refinado, devido às diversas entidades que poderão interagir com os dados. Desse modo, os direitos de acesso aos recursos precisam ser garantidos aos usuários somente pelo tempo necessário. Isto evita que erros de vazamento de informações confidenciais para sujeitos não autorizados ocorram, protegendo os usuários de falhas não intencionais. Além disso, as políticas de acesso devem fornecer suporte à realização das tarefas de trabalho dos indivíduos.

Um dos fatores importantes de controle de acesso é o locatário/inquilino, que é um cliente, como um hospital, uma clínica ou farmácia do sistema de saúde. Cada locatário possui diversos usuários, que podem ser pacientes, médicos, enfermeiros e técnicos. Uma interface de usuário trata-se de um funcionário de um locatário ou de um paciente do sistema de saúde. Continuando com os fatores importantes, uma tarefa é uma unidade imprescindível da atividade empresarial, as quais são associadas aos usuários a partir da função que possuem de modo que, seus direitos de acesso são concedidos para cumprir as tarefas atribuídas. Os recursos são objetos de controle de acesso, como banco de dados e arquivos. Já a função de negócios é atribuída a cada usuário baseado nas atividades de negócio que executam na organização, concedendo o acesso aos recursos necessários. A permissão é uma concessão para realizar uma operação em um objeto, enquanto a sessão mapeia um usuário para funções diferentes. O fluxo de trabalho trata-se de um conjunto de tarefas empregadas na execução de uma função de negócios de modo que, as tarefas que estão inclusas no fluxo de trabalho demandam controle de acesso ativo e as que não pertencem requerem um controle de acesso passivo. Portanto, um sistema de saúde possui tarefas de fluxo de trabalho, como gravar uma prescrição médica, e tarefas que não são do fluxo de trabalho, como listar os pacientes atuais. A execução das tarefas de fluxo de

trabalho tem uma ordem a ser seguida e estão disponíveis por um período de tempo específico.

As regras de negócios são fundamentos importantes e trata-se das práticas padrão exercidas pelos usuários que a organização segue e incluem: Privilégio mínimo, Separação de deveres, Delegação de Tarefas, Restrições espaciais e temporais e Classificação de tarefas. A Classificação de tarefas apresenta quatro possíveis classes, que são Classe privada, Supervisão de classe, Fluxo de trabalho da classe e Aprovação da classe. Outro ponto importante é o Escopo, pois o controle de acesso é gerenciado neste nível. Cada escopo herda funções, permissões e regras de negócio de um escopo pai a partir da estratégia de relacionamento estabelecida pelo sistema de saúde. Com base nestes itens listados, temos que as diferentes organizações de saúde acessam diferentes instâncias do sistema a partir de uma base de dados centralizada e acessível por meio da nuvem.

Ao abordar o controle de acesso baseado em função de tarefa com restrições, destaca-se pontos importantes deste modelo, como as relações existentes entre as funções, usuários, tarefas de fluxo de trabalho e permissões, e a utilização de funções para promover suporte ao controle de acesso passivo e as tarefas para fornecer suporte ao controle de acesso ativo. São elencados os possíveis pares de atribuições de entidade deste modelo, os quais são: atribuição de inquilino-usuário (um sistema em nuvem possui diversos inquilinos, e cada um deles com vários usuários); funções de usuário como assinatura (usuários e funções possuem um relacionamento muitos para muitos, em que um usuário pode ser associado a uma ou mais funções e uma função pode ser associada a mais de um usuário); atribuição de tarefa-função (tarefas e funções também possuem um relacionamento muitos para muitos de modo que, uma função pode ser associada a diversas tarefas e uma tarefa pode ser associada a mais de uma função); atribuição de permissão de tarefa (concessão de permissão para a execução da tarefa); e atribuição de fluxo de trabalho de tarefa (para ser atribuída a um fluxo de trabalho, a tarefa deve pertencer a classes de fluxo de trabalho ou aprovação).

As restrições de tarefa são outro aspecto fundamental nas regras de negócio. A primeira delas é o privilégio mínimo, que é obtido através de instâncias da tarefa: a permissão de acesso se inicia junto com a tarefa e as permissões de controle de acesso são revogadas quando a tarefa é finalizada. Isto promove menos privilégio e o controle de acesso refinado. A segunda é a separação estática e dinâmica de tarefas. A separação estática é feita no nível de definição de tarefa, coibindo a atribuição de mais de uma tarefa

mútua para a mesma função em um mesmo tempo. A separação dinâmica de tarefas é aplicada no nível das instâncias de tarefa, impedindo que duas ou mais tarefas exclusivas sejam executadas pela mesma função. A terceira é a delegação, que é realizada através da concessão de tarefa refinada pelo usuário atribuído inicialmente. Por fim, tem-se as restrições espaço temporais, em que a localização e a hora do usuário são consideradas ao se conceder acesso a uma tarefa.

Narayanan [49] descreve a implementação proposta, a qual implanta o mecanismo de controle de acesso baseado em função de tarefa no Amazon EC2 e aplicações em Java. As informações de controle de acesso baseado em tarefa, assim como a aplicação do sistema de saúde, são armazenadas em um banco de dados SQL. Nesse modelo, o administrador do sistema de locatário cria uma função de administrador para cada locatário, concedendo direitos de acesso a eles para gerenciarem a autenticação e autorização referentes ao seu próprio domínio. Políticas flexíveis são criadas e as restrições de acesso são associadas aos usuários e tarefas, para que o uso inapropriado de permissão seja coibido. Quando um usuário realiza a autenticação, o sistema verifica as credenciais do mesmo e estabelece suas funções. A partir das funções, as tarefas para as funções ativas são exibidas pela página de seleção de tarefas. Para um usuário administrador, tarefas de gerenciamento envolvendo manutenção de usuário, funções e fluxo de trabalho também são provisionadas. Todos os procedimentos aplicados nesta fase visam proporcionar segurança contra invasores.

3.3 Aspectos Relevantes dos Trabalhos Relacionados para a Proposta

As regras de construção das políticas de permissionamento para aplicações e recursos na nuvem diferem quando se trata de agentes pessoas físicas ou aplicações, como apresentado por Sanders [45]. Dessa forma, esta seção será dividida entre aspectos relevantes para permissionamento de aplicações e de usuários.

3.3.1 Permissionamento de Aplicações

O comportamento das aplicações é mais previsível do que os comportamentos dos usuários, pois as necessidades de acesso não mudam com tanta frequência. Isso indica que as políticas de controle de acesso geradas não precisam ser recicladas com muita frequência, permitindo tempos de operação maiores.

Para as aplicações, existe um momento importante em que as necessidades de permissionamento podem ser alteradas, que consiste no processo de implantação de uma nova versão da aplicação. Nas situações de implantação de novas versões, é interessante mapear os relacionamentos entre as tarefas realizadas e os recursos envolvidos, assim como o relacionamento entre as entidades envolvidas na execução das funções e os recursos utilizados por cada uma. Com isso, é possível ter uma visão mais clara do que realmente é necessário para executar com êxito uma determinada função, evitando a deliberação excessiva de privilégios. Tal ação pode ser realizada através da análise do código-fonte em busca das chamadas do Kit de Desenvolvimento de Software (*Software Development Kit - SDK)* do provedor de nuvem. Durante a construção das aplicações, os desenvolvedores definem as chamadas de aplicação que vão utilizar e se utilizam de definições de políticas necessárias para que a aplicação funcione. Gill et al. [39] fez uma análise de chamadas de API na AWS e mostrou o quanto as políticas de aplicações criadas possuem privilégios excessivos.

As aplicações, por sua vez, se dividem em dois tipos de ações: as ações agendadas e as ações provocadas por usuários. As ações agendadas são aquelas em que rotinas são realizadas com base em procedimentos pré-configurados e constantes, como a capacidade de ligar e desligar um servidor com base no horário agendado. Enquanto as ações provocadas por usuários são aquelas que derivam de ações dos usuários na aplicação, como o carregamento de um arquivo a partir de um repositório de objetos. Para ambos os casos, o modelo RBAC com a especialização de tarefas pode ser considerado.

No cenário das ações agendadas, deve-se ativar e desativar as permissões de execução da aplicação com base no procedimento a ser executado a cada momento, garantindo que uma execução errônea do procedimento em horário não regular seja negado.

Por outro lado, as ações provocadas por usuários são mais complexas. Dependendo do sistema e da sensibilidade dos dados, é importante considerar a duração da concessão de privilégios, de modo que os profissionais que acessam as informações provisionadas pelo sistema o façam somente nos horários dos respectivos expedientes. Além disso, é

importante que a aplicação registre nos logs do provedor de nuvem quem foi o usuário que executou a ação além da informação da função utilizada pela aplicação. Isso promove a segurança dos dados e evita inconsistência nas informações. Para que seja possível empregar um controle de acesso que delegue concessões com base no tempo, pode ser empregado o controle de acesso baseado em funções de tarefas. Com ele, é possível especificar em quais condições das regras de negócio da aplicação, um determinado usuário pode acessar um conjunto de dados ou executar um serviço. Isso permite ter um sistema dinâmico, que adapta as políticas para o cenário atual de cada funcionário.

Mesmo que não haja dinâmica nos acessos permitidos, a solução proposta deve ser capaz de fazer atualizações de modo a aprender à medida que os usuários utilizam os sistemas e esses sistemas fazem uso das APIs do provedor de nuvem. A solução também deve ser capaz de lidar com exceções, concedendo privilégios que não foram previstos na fase de implementação. Para dar a escala de privilégios adequada, é importante considerar também a divisão de tarefas e funções com base na hierarquia. Isso permite ter um controle mais preciso, de modo que as operações sobre os dados (leitura, escrita, inserção, deleção) só sejam permitidas se o nível de acesso no momento da ação o permitir.

3.3.2 Permissionamento de Usuários

Ações realizadas diretamente por usuários nos serviços do provedor de nuvem apresentam maior complexidade em relação às aplicações, pois as ações que um usuário realiza possuem as mais variadas naturezas, inclusive aquela exploratória, em que o usuário deseja aprender sobre novos serviços do provedor de nuvem. Portanto, o comportamento dos usuários de uma organização não é tão previsível quanto o das aplicações. Sobretudo, a longo prazo, os usuários podem mudar de cargo ou saírem da empresa. Isso implica uma concessão de privilégios mais dinâmica, que seja atualizada com uma frequência maior que a das aplicações e sem um evento característico como um a implantação de uma nova versão da aplicação. Quanto maior a organização, mais pessoas envolvidas e maior é a dinâmica das funções desempenhadas pelos funcionários. Assim, é importante ter um mapeamento das relações entre as entidades, as funções e os recursos, para que não haja excesso de privilégios e nem privilégios insuficientes.

Uma estratégia a se considerar, é fazer uma análise a fim de agrupar os usuários por funções executadas. Assim, é possível gerenciar melhor as políticas, sendo mais assertivo

ao concedê-las, e as mudanças são implementadas com mais facilidade. Além disso, é importante identificar atributos do ambiente e do usuário, de forma a especificar as ações de acordo com esses atributos. Por exemplo, em um ambiente de desenvolvimento, o usuário deve conseguir realizar ações exploratórias em novos serviços, o que não deve ser permitido no ambiente de produção. Para tanto, logs de acesso e do histórico de configurações dos recursos da nuvem são úteis para acompanhar as mudanças que devem ser acrescentadas à solução. Além disso, no acesso de usuários, uma preocupação adicional que se deve ter é a capacidade de escalar privilégios, em que um usuário possui permissão de mudar as suas permissões ou de criar usuários com permissões maiores, conforme descrito e analisado em Hu et al. [51].

A solução proposta deve continuar aprendendo com o ambiente, pois a dinâmica das organizações deve ser atendida pelo mesmo pois, do contrário, a solução deixará de ser eficaz tornando-se obsoleta. O acompanhamento dos logs coopera para a evolução das políticas de acesso criadas, que apresentam melhor desempenho se forem geradas automaticamente, pois ajudam a identificar as mudanças no ambiente e a mitigar o subprivilégio e o superprivilégio. Com isso, ter políticas de controle de acesso geradas automaticamente evitam erros de permissão, principalmente para grandes organizações. A solução deve ser capaz de aderir às exceções que possam surgir após a implementação, demandando um tratamento mais específico para estes casos. A solução para o caso dos usuários considera as funções e os atributos para realização de regras mais assertivas.

Capítulo 4 - Proposta de Criação de Base de Dados para Análise de Privilégios em Nuvem

A correta atribuição de permissões para os agentes da nuvem é um desafio em aberto. Este trabalho visa identificar oportunidades de melhorias nas propostas existentes, através de uma proposta que garanta o privilégio mínimo em uma nuvem pública considerando o contexto de cada cliente da nuvem e suas necessidades únicas. Neste Capítulo são apresentados elementos da base de dados reais de ambientes de produção e se divide em três seções: A Seção 4.1 apresenta os serviços da Amazon Web Services (AWS) cujas informações serão coletadas para a base de dados criada; A Seção 4.2 descreve como os dados serão coletados; e a Seção 4.3 discute sobre como os dados coletados são relevantes para a construção de políticas para aplicações e usuários.

Conforme destacado em Lei e Tripunitara [52], o problema de aprendizado de políticas de controle de acessos é difícil e este trabalho visa a construção de bases de dados reais. Este trabalho usa dados de auditoria de contas AWS de ambientes de produção de clientes reais da AWS, que são geridas pela empresa e-Core[16], juntamente com as políticas de acesso vigentes em cada um dos ambientes para todos os sujeitos (pessoas e aplicações) da conta. Em todo ambiente em nuvem, pode-se ter dois tipos diferentes de acessos: (i) usuários humanos que configuram e gerenciam o ambiente e (ii) acessos de aplicações que precisam de acessos às APIs da nuvem para executarem a função que foram designadas. Este trabalho trata de ambos os tipos de acesso trazendo abordagens adaptadas para cada um deles. Os dados de auditoria obtidos envolvem registro de logs das chamadas de APIs aos serviços da nuvem de todo o ambiente e registro de alterações de todos os recursos existentes na conta AWS.

A Figura 5.1 apresenta um diagrama com o fluxograma de decisão da proposta de base de dados, que considera as fontes de informação a serem coletadas e a sequência de processos que são feitos no fluxo da solução, sendo uma visão macro da solução. Como a contribuição deste trabalho é a construção da base, define-se uma base conceitual sobre como esses dados devem ser consumidos por um algoritmo genérico que possam ter aplicação prática em ambientes de produção. Após a geração das políticas, um validador das políticas é criado para garantir que o acesso será suficiente. A partir dos registros do AWS Config é possível identificar os serviços AWS utilizados no ambiente e simplificar a

[16] Empresa que adquiriu a empresa Solvimm, fundada pelo autor do trabalho.

construção por tecnologias. A partir das Políticas de Permissionamento é possível obter a relação de permissionamento configurados para cada *Principal,* que é a pessoa ou aplicação que pode executar uma requisição para uma ação ou operação em um recurso AWS, ao longo do tempo. Por fim, a partir dos registros dos eventos de auditoria do *Amazon CloudTrail*, são identificados quais ações em serviços são utilizadas por cada *Principal*. Com base nessas três informações principais, que constituem a base de dados construída, futuros algoritmos podem gerar Políticas de Permissionamento para usuários e aplicações. Essas políticas são avaliadas através da simulação de execuções de ações que os usuários e funções devem executar. Após que a Política de Permissão seja atribuída para o *Principal*, dado que a trilha de auditoria do *Amazon CloudTrail* continua a ser coletada, uma eventual recusa de acesso retroalimentará a base de dados, apoiando a reformulação da Política de Permissionamento dado o ambiente dinâmico. A base de dados construída, pela sua natureza online, pode possibilitar diversas abordagens de aprendizados para algoritmos a serem propostos em trabalhos futuros.

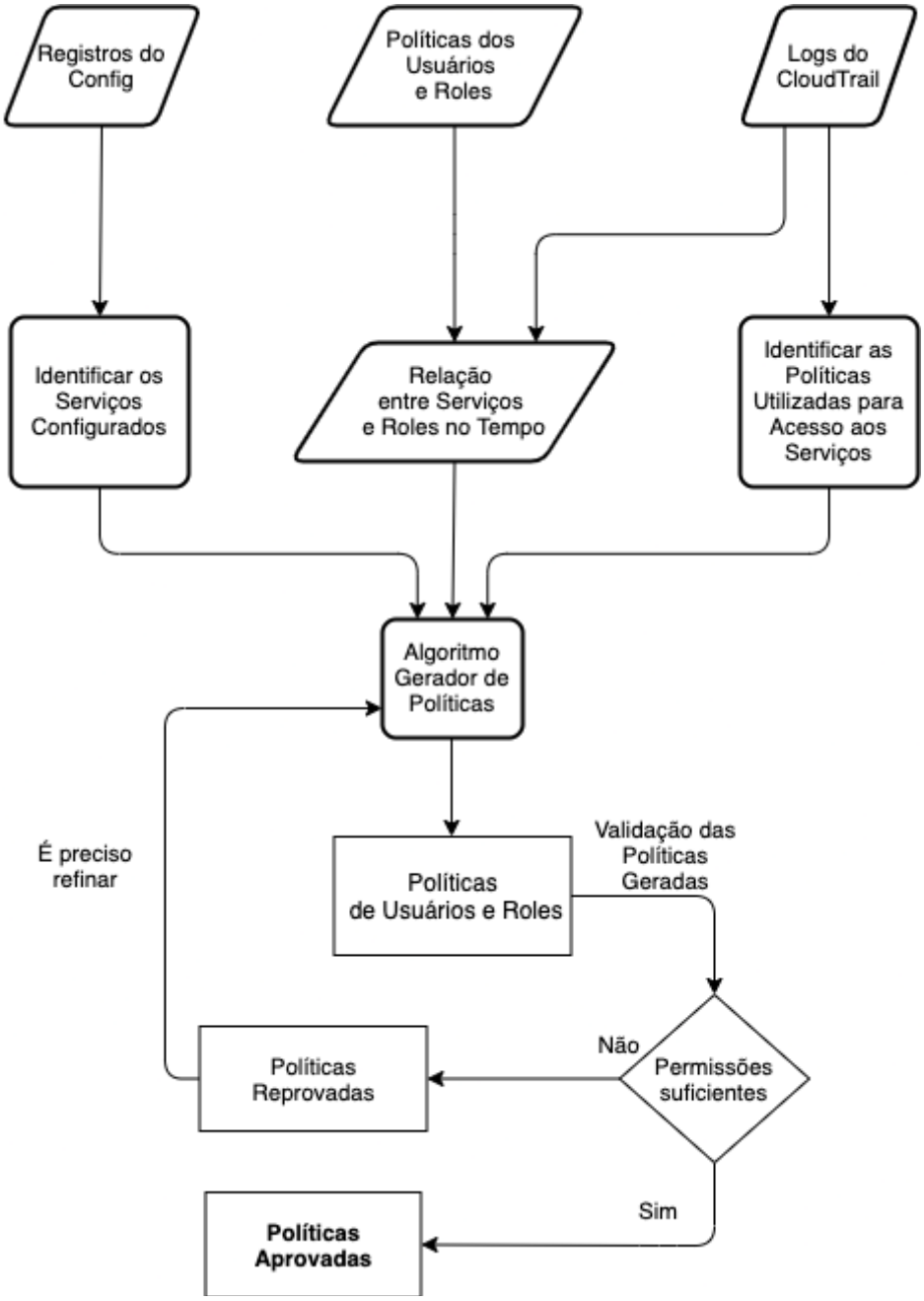


Figura 5.1- Fluxograma de Construção e Consumo da Base de Dados Criada

4.1 Obtenção dos Dados para Análise

Os dados serão coletados dos ambientes de produção através de extratores de dados a partir de arquivos de logs ou das configurações dos serviços. São, basicamente três fontes de dados de cada conta AWS analisada que serão consolidados em um repositório de dados (*datalake*) no Amazon S3 da conta AWS deste projeto. Esta seção visa descrever cada uma das fontes de dados e o processo de coleta realizado. As políticas vigentes são obtidas

diretamente do serviço de gerenciamento de identidades e acessos, o AWS IAM (Seção 4.1.1), enquanto os logs de acessos das APIs dos serviços de gerenciamento da nuvem são obtidos pelo serviço Amazon CloudTrail (Seção 4.1.2) e, por fim, os registros de configuração histórica dos recursos da conta AWS são armazenados pelo AWS Config (Seção 4.1.3).

4.1.1 Serviço AWS IAM

O serviço AWS Identity and Access Management (IAM)[17] é um serviço focado no controle de acesso aos recursos de TI na AWS através das chamadas de API dos serviços web da AWS. Através do IAM, pode-se controlar quem está autenticado e autorizado para usar esses recursos, tanto para ações de leitura quanto para ações de escrita e gerenciamento. Uma conta AWS é contêiner para os recursos criados e gerenciados na AWS com faturamento exclusivo para um cliente da AWS, que é a empresa que contrata os serviços da nuvem. No momento que uma conta AWS é criada por algum cliente, este possui um acesso principal, conhecido como Root Access. Uma das primeiras ações que devem ser seguidas é a criação de um usuário no IAM e o armazenamento seguro das credenciais do Root, cujo uso deve ser extremamente restrito. A partir desse ponto, toda a gestão de usuários, grupo, funções e permissões é feita através do serviço AWS IAM. No contexto do gerenciamento de identidades e acessos, o cliente é a empresa que contrata os serviços da AWS. Os usuários são os colaboradores da empresa que criam e gerenciam os recursos na nuvem, que são os servidores, bancos de dados, espaços de armazenamento, entre outros.

O serviço provê toda a infraestrutura necessária para controlar a autenticação de usuários e recursos e implementar controle de autorização na conta AWS, sendo compatível tanto com o modelo de *Role-based Access Control (RBAC)* quanto com o modelo *Attibute-based Access Control* (ABAC). Dentre as estruturas existentes no IAM, abaixo apresentam-se os principais elementos que compõe o IAM, como demonstrado na Figura 5.2[18]. Em resumo, os Recursos do IAM (usuários, grupos, funções, políticas de permissões e provedor de identidade) são armazenados dentro da base de dados do IAM.

[17] Página do serviço na AWS: https://aws.amazon.com/pt/iam/
[18] Extraído da Documentação Oficial: https://docs.aws.amazon.com/IAM/latest/UserGuide/introstructure.html

Com base nesses recursos do IAM, criam-se relacionamentos entre eles para atribuir autorização a uma pessoa ou aplicação (IAM *Principals*) que usa a conta AWS para manipular os recursos na AWS, como um servidor virtual.

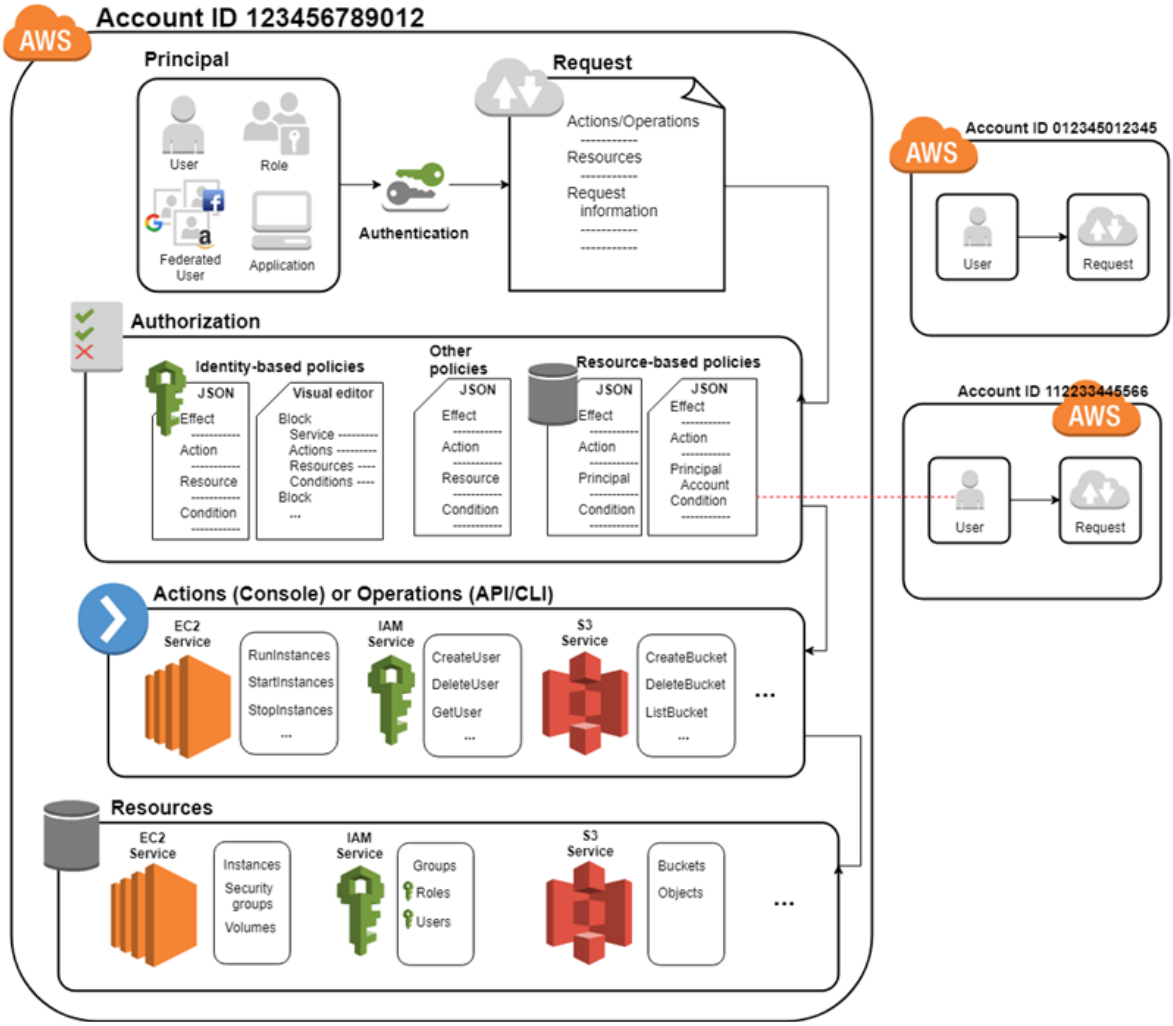


Figura 5.2 - Estrutura dos principais elementos do serviço AWS IAM, extraído da Documentação Oficial da AWS, com a relação entre os elementos. Fonte: Amazon AWS[18].

Um *Principal* está autenticado na conta da AWS, sendo o próprio usuário *root*, que não deve ser utilizado, ou uma entidade do IAM através dos usuários ou funções (*Roles*). Quando a estrutura de autenticação é federada com outros diretórios de usuários, como *Active Directory*, um LDAP ou mesmo AWS IAM de outra conta AWS, a entidade que representa o acesso é uma função do IAM (IAM Roles). A federação de identidades ocorre

quando um ambiente na AWS autentica um usuário com base em um diretório de usuários externos. Esse processo pode ocorrer tanto para acessos de usuários quanto de aplicações. Em ambos os casos, o registro será feito considerando toda a cadeia de autenticação e registrando o usuário final, o que permite que a política seja definida de acordo com atributos do usuário e não apenas com o acesso assumindo ao conectar na conta AWS.

Cada ação que o *Principal* executa no ambiente da AWS, seja através do console WEB de gerenciamento, linha de comando ou por APIs, é realizada através de requisições HTTP para o serviço AWS desejado. Essas requisições, antes de serem processadas pelo serviço solicitado, são avaliadas pelo AWS IAM, que avalia se o *Principal* tem autorização para realizar a ação solicitada ao serviço AWS. As requisições possuem as informações para a completa avaliação, consistindo nos elementos:

a) **Principal** é a pessoa ou aplicação usada para enviar a requisição, como, por exemplo, um usuário ou uma função (*role)* associada. Importante ressaltar que os dados do Principal incluem as informações que remetem a entidade que realizou o login, permitindo o rastreamento da entidade que se autenticou inicialmente em caso de federação de entidades, por exemplo. Com base na informação do *Principal*, o AWS IAM obtém as políticas de permissionamento associadas a esta entidade.
b) **Ações ou Operações** que o *Principal* realiza na AWS, tais como, criar uma máquina virtual ou remover um arquivo armazenado.
c) **Recursos** são os identificadores do objeto da AWS sobre o qual a ação será executada, como, por exemplo, um banco de dados ou uma fila de mensagens.
d) **Dados do Ambiente** são informações sobre o contexto da requisição que incluem o endereço IP da chamada, a aplicação usada para fazer a requisição (*user agent do protocolo HTTP*), configurações do protocolo SSL e o timestamp da requisição.
e) **Dados do Recurso** são relacionados ao recurso sendo requisitado como o nome de uma tabela de base de dados ou o rótulo (*tag*) de um servidor virtual.
f) **Políticas de Permissões** compreendem um documento que armazena, em formato JSON, as regras de permissões a serem associadas aos *Principals* sobre os Recursos.

O AWS IAM avalia se o Principal tem permissão ou não para executar aquela ação dentro do contexto solicitado. A decisão de autorização é então passada ao serviço que executa somente em caso de autorização. Para este trabalho, as configurações de Principal e suas Políticas de Permissões associadas serão utilizadas para o conhecimento das permissões existentes em cada principal ao longo do tempo.

4.1.2 Serviço Amazon CloudTrail

Todas as requisições realizadas por serviços na AWS juntamente com a avaliação de autorização do AWS IAM é armazenada em uma trilha de auditoria no Amazon CloudTrail[19]. A trilha de auditoria é uma sequência ordenada no tempo de todas as chamadas de uma determinada conta AWS. O serviço possibilita a governança, conformidade, auditoria operacional e auditoria de riscos em uma conta AWS. Cada requisição é armazenada na trilha de auditoria como um evento. Este serviço é fundamental para a segurança das contas AWS e já vem ativado por padrão em todas as contas, armazenando os dados por 90 dias. Para que seja possível o armazenamento por período maior, uma boa prática realizada pelas empresas é o armazenamento permanente destes registros de log através de arquivos em um espaço de armazenamento no serviço Amazon S3. Quanto ativo é uma Conta AWS, o CloudTrail recebe os dados de logs de todos os serviços em todas as regiões habilitadas na Conta AWS, sendo um repositório global de eventos de auditoria.

Para este trabalho, os logs de auditoria serão utilizados para ter o rastreamento de cada uma das requisições realizadas por cada Principal em uma janela de tempo de forma a entender as ações necessárias e desejáveis que um Principal possa realizar. Importante destacar que este trabalho não se baseia apenas no histórico de utilização, como apresentado nos trabalhos relacionados [45, 46, 47], pois em uma utilização real, é comum as configurações estarem muito permissivas e os usuários realizarem chamadas que não deveriam realizar, o que faz as políticas geradas automaticamente com base apenas no histórico de requisições passados ser muito permissivo.

[19] Página do serviço da AWS: https://aws.amazon.com/pt/cloudtrail/

4.1.3 Serviço AWS Config

Além dos registros das requisições nos eventos do Amazon CloudTrail, este trabalho também utiliza os registros do *AWS Config*[20]. Os itens de configuração do *AWS Config* possuem o histórico do estado detalhado de configuração de cada recurso criado na conta AWS. Com isso, permite identificar quais os serviços são utilizados pelo cliente na contadesde a sua criação e como esses recursos se relacionam com outros. Dessa forma, é possível definir os agrupamentos de recursos e serviços utilizados pelos clientes nas contas AWS, como, por exemplo, para um servidor virtual, o *AWS Config* o relaciona com os discos, interfaces de rede, métricas de monitoramento, endereços IPs e outros recursos relacionados a este servidor.

Com isso, neste trabalho, esses dados são utilizados para identificar os serviços utilizados pelos clientes de forma a permitir o agrupamento dos recursos por tecnologias, possibilitando uma dimensão simplificada para a construção das políticas de permissionamento com base nesses serviços e grupos. Tal informação permite também a criação de correlações entre os serviços que podem gerar conhecimento a ser utilizado entre múltiplos ambientes de clientes diferentes. Por exemplo, um desenvolvedor deve ter permissionamento diferente caso o ambiente esteja configurado em servidores virtuais, containers ou funções no ambiente em nuvem.

4.2 Processo de Coleta de Dados

A análise desse trabalho será realizada utilizando dados reais de clientes da AWS cujos ambientes na AWS são gerenciados pela e-Core. Os dados de cada uma das contas AWS dos clientes são coletados das contas AWS de cada cliente através de um mecanismo de coleta implementado neste trabalho. O mecanismo funciona através de ferramentas de ETL (*Extract, Tranform, Load*) executadas diariamente para coletar de cada um dos ambientes os conjuntos de dados listados:

a) configuração dos *Principals* (usuários e funções) das contas com suas respectivas políticas de permissionamento, através do AWS IAM;

[20] Página do serviço da AWS: https://aws.amazon.com/pt/config/

b) registros de auditoria com as requisições realizadas na conta AWS, através do Amazon CloudTrail; e
c) registros de histórico de configuração dos recursos AWS e seus relacionamentos com outros recursos, através do AWS Config.

Todos esses dados são armazenados em um repositório de dados (*datalake*) no Amazon S3, em formato de arquivos JSON. A solução funciona através de uma aplicação criada que realiza acesso às contas AWS dos clientes utilizando Roles de acesso. Essas roles possuem permissão de leitura para coletar os dados necessários. Esse procedimento é o procedimento comum realizado em ambientes AWS para aplicações acessarem serviços em outras contas. Importante mencionar que os acessos diários da coleta irão também compor os dados e essa aplicação é uma das aplicações que será identificada nos dados.

Foram coletados dados de ambientes de produção de 20 empresas que utilizam a AWS para executar suas operações diárias. Destas, 10 empresas utilizam serviços de máquina virtual, compreendendo servidores, bancos de dados, redes, monitoramento e backup. Outras 5 empresas possuem ambientes em clusteres de containers, que compreendem serviços diferentes da AWS e demanda permissionamento diferentes. E as últimas 5 empresas possuem aplicações em arquitetura Serverless, com serviços nativos da nuvem para suas aplicações. Em todos os casos, existem usuários que acessam o ambiente para gerenciá-lo e aplicações que acessam serviços e precisam de permissionamento para acesso a dados e chamadas de serviços.

4.3 Discussão sobre a Relevância dos Dados para a Construção de Algoritmos de Geração de Políticas

Com a base de dados construída, trabalhos futuros podem estudar diferentes algoritmos de geração de políticas de permissionamento. Este trabalho apresenta uma visão geral dos elementos que serão necessários construir e avaliar para a construção algoritmos realmente úteis em produção, o que foi considerado na construção da base de dados proposta. Desse modo, segundo o diagrama da Figura 4.1, inicialmente os registros do *AWS Config* são usados para determinar os serviços em uso na conta, os registros de logs do *Amazon CloudTrail* para identificar quais as políticas utilizadas pelos *Principals* (usuários e funções) e, a partir desses conjuntos de dados, obter a relação entre os serviços cadastrados e as funções necessárias à sua utilização, determinando, inclusive, se o

Principal se refere a um usuário de pessoa ou a uma aplicação. Os comportamentos dos usuários possuem uma tendência maior de alteração se comparado as aplicações, pois o ambiente corporativo é dinâmico, principalmente para grandes corporações.

Ainda fazendo uso dos registros de logs do *Amazon CloudTrail* e dos registros históricos do *AWS Config*, para que sejam geradas políticas de privilégio mínimo, deve-se atrelar aos *Principals* após a validação. O resultado do algoritmo será Políticas de Permissionamento, que precisam ser validadas para que possam garantir que atendem à realidade do cliente através dos dados do *AWS Config*. Caso não esteja de acordo, será necessário refiná-la. Devido à característica evolutiva de um ambiente real na AWS, que passa a adotar novos serviços e tem certa dinâmica na existência dos *Principals* com a entrada e saída de novos usuários além da criação e remoção de aplicações, a abordagem requer algoritmos que constantemente possam revalidar a as políticas dos *Principals* das contas AWS, pois é importante considerar o ajuste das políticas com o passar do tempo.

Capítulo 5 - Conclusão

A computação em nuvem já é realidade no mercado e tem sido largamente adotada por grandes empresas. Com o aumento da complexidade dos ambientes na nuvem e a concentração de uso de serviços hospedados fora das fronteiras da empresa, surgem desafios de segurança que precisam ser endereçados. Este trabalho apresentou duas propostas para melhorias de segurança no uso da nuvem por empresas, sobretudo grandes empresas e gerou 2 publicações em revistas internacionais [23, 53] e 2 publicações em congressos [35, 54] considerando ambas as propostas.

A primeira proposta focou nos desafios de classificação e segregação do tráfego de rede em ambientes de alta escala, uma vez que os serviços em nuvem fazem com que o perfil de uso das redes seja uma agregação dos tráfegos internos para a conexão com a Internet. O algoritmo proposto foi baseado em cadeias de Markov aplicadas a 5 perfis de uso obtidos através de algoritmo de clusterização dos fluxos da rede. Com esses perfis, pode-se determinar a carga necessária na rede e o quanto os equipamentos internos podem ser economizados. Nos testes com a rede da Universidade Federal Fluminense, identificou-se o perfil do tráfego e previu-se a carga nos pontos de acesso da rede em mais de 90% dos cenários testados.

A avaliação da estratégia proposta, segundo estudo, demonstrou que sua assertividade é maior quando há maior tráfego na rede, pois há maior massa de dados para a análise. Com poucos dados, sua margem de erro aumenta, uma vez que existe menos informação para previsão. A proposta incluiu ainda a análise de perfis de uso através de um algoritmo de aprendizado de máquina não supervisionado *(k-means*), que tem seu melhor desempenho com 5 grupos de perfis de uso. Esses perfis de uso são combinados à previsão de acessos em cada ponto de acesso. Portanto, através da estratégia proposta é possível estimar a carga de uso em cada ponto de acesso da rede por perfil, possibilitando a construção de soluções de detecção de anomalia de redes, planejamento e economia de energia através do desligamento de dispositivos ou ainda na implantação de soluções inteligentes de cache de dados e controle de qualidade de serviço.

A segunda proposta consistiu em uma análise de propostas e na construção de uma base de dados para análise de privilégios de acessos para garantir o privilégio mínimo no uso dos ambientes em nuvem. A proposta se diferencia das analisadas nos trabalhos

relacionados por correlacionar políticas com os serviços em utilização e comparar esses dados entre clientes distintos da nuvem. A coleta de dados para a análise do algoritmo se dá a partir de ambientes de produção na nuvem da AWS com a consolidação em um ambiente de um repositório de dados (*datalake*) do projeto. São necessários futuros estudos para implementação do algoritmo de geração de políticas e avaliação com os dados coletados no âmbito desta proposta.

Referências

[1] FOSTER, I.; ZHAO, Y.; RAICU, I.; LU, S. Cloud computing and grid computing 360-degree compared. In: IEEE. 2008 grid computing environments workshop. [S.l.], 2008. p. 1–10.

[2] SHAHZAD, F. State-of-the-art survey on cloud computing security challenges, approaches and solutions. Procedia Computer Science, Elsevier, v. 37, p. 357–362, 2014.

[3] DILLON, T.; WU, C.; CHANG, E. Cloud computing: issues and challenges. In: IEEE.2010 24th IEEE international conference on advanced information networking and applications. [S.l.], 2010. p. 27–33.

[4] LI, M.; YU, S.; REN, K.; LOU, W.; HOU, Y. T. Toward privacy-assured and searchable cloud data storage services. IEEE Network, IEEE, v. 27, n. 4, p. 56–62, 2013.

[5] IANKOULOVA, I.; DANEVA, M. Cloud computing security requirements: A systematic review. In: IEEE. 2012 Sixth International Conference on Research Challenges in Information Science (RCIS). [S.l.], 2012. p. 1–7.

[6] CISCO, V. N. I. Global mobile data traffic forecast update, 2017–2022 white paper.Document ID, v. 1486680503328360, 2019.

[7] BISWAS, S.; BICKET, J.; WONG, E.; MUSALOIU-E, R.; BHARTIA, A.;AGUAYO, D. Large-scale measurements of wireless network behavior. In: Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication. London, United Kingdom: ACM, 2015. (SIGCOMM ’15), p. 153–165. ISBN 978-1-4503-3542-3. Disponível em: <http://doi.acm.org/10.1145/2785956.2787489>.

[8] BALBI, H.; FERNANDES, N.; SOUZA, F.; CARRANO, R.; ALBUQUERQUE,C.; MUCHALUAT-SAADE, D.; MAGALHãES, L. Centralized channel allocation algorithm for ieee 802.11 networks. In: 2012 Global Information Infrastructure and Networking Symposium (GIIS). [S.l.: s.n.], 2012. p. 1–7. ISSN 2150-3281.

[9] MATTOS, Diogo Menezes Ferrazani; MEDEIROS, Dianne Scherly Varela de; FERNANDES, Natalia ; MAGALHAES, Luiz . Uma Abordagem Não Supervisionada para Inferir Qualidade de Experiência em Redes Sem Fio de Grande Escala. In: WORKSHOP DE GERÊNCIA E OPERAÇÃO DE REDES E SERVIÇOS (WGRS), 24. , 2019, Gramado. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2019 . p. 57-70. ISSN 2595-2722. DOI: https://doi.org/10.5753/wgrs.2019.7683

[10] GHOSH, A.; JANA, R.; RAMASWAMI, V.; ROWLAND, J.;

SHANKARANARAYANAN, N. K. Modeling and characterization of large-scale wi-fi traffic in public hot-spots. In: 2011 Proceedings IEEE INFOCOM. [S.l.: s.n.], 2011. p. 2921–2929. ISSN 0743-166X.

[11] QIAN, F.; WANG, Z.; GERBER, A.; MAO, Z.; SEN, S.; SPATSCHECK, O. Profiling resource usage for mobile applications: A cross-layer approach. In: Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services. Bethesda, Maryland, USA: ACM, 2011. (MobiSys '11), p. 321–334. Disponível em: <http://doi.acm.org/10.1145/1999995.2000026>.

[12] SHYE, A.; SCHOLBROCK, B.; MEMIK, G.; DINDA, P. A. Characterizing andmodeling user activity on smartphones: summary. In: ACM. ACM SIGMETRICS Performance Evaluation Review. [S.l.], 2010. v. 38, p. 375–376.

[13] OLIVEIRA, L.; OBRACZKA, K.; RODRÍGUEZ, A. Characterizing user activity inwifi networks: University campus and urban area case studies. In: Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems. Malta: ACM, 2016. (MSWiM '16), p. 190–194. ISBN 978-1-4503-4502-6.

[14] MAGALHãES, L. C. S.; MATTOS, D. M. F. Caracterização do uso de uma redesem fio de grande porte distribuída por uma ampla área. In: SBC. 17º Workshop emDesempenho de Sistemas Computacionais e de Comunicação (WPerformance 2018).[S.l.], 2018. v. 17.

[15] LOPEZ, M. A.; MATTOS, D. M.; DUARTE, O. C. M.; PUJOLLE, G. A fast unsupervised preprocessing method for network monitoring. Annals of Telecommunications, Springer, p. 1–17, 2018.

[16] BOUTABA, R.; SALAHUDDIN, M. A.; LIMAM, N.; AYOUBI, S.; SHAHRIAR,N.; ESTRADA-SOLANO, F.; CAICEDO, O. M. A comprehensive survey on machinelearning for networking: evolution, applications and research opportunities. Journal of Internet Services and Applications, v. 9, n. 1, p. 16, Jun 2018. Disponível em:<https://doi.org/10.1186/s13174-018-0087-2>

[17] MATTOS, D. M. F.; MEDEIROS, D. S. V. de; FERNANDES, N. C.; MAGALHãES,L. C. S. Uma abordagem não supervisionada para inferir qualidade de experiência em redes sem fio de grande escala. In: Workshop de Gerência e Operação de Redes e Serviços do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos - WGRS'2019. [S.l.: s.n.], 2019.

[18] LYU, F.; REN, J.; CHENG, N.; YANG, P.; LI, M.; ZHANG, Y.; SHEN, X. S. Bigdata analytics for user association characterization in large-scale wifi system. In: IEEE ICC 2019 - Empowering Intelligent Communications (ICC'19). Shanghai, China: [s.n.],2019.

[19] MATTOS, D. M.; DUARTE, O. C. M.; PUJOLLE, G. Profiling software definednetworks for dynamic distributed-controller provisioning. In: IEEE. 2016 7th International Conference on the Network of the Future (NOF). [S.l.], 2016. p. 1–5.

[20] M. B. d. A. Rodrigues et al., "An Efficient Strategy with High Availability for Dynamic Provisioning of Access Points in Large-Scale Wireless Networks," 2022 5th Conference on Cloud and Internet of Things (CIoT), Marrakech, Morocco, 2022, pp. 92-99, doi: 10.1109/CIoT53061.2022.9766688.

[21] CLAISE, B. Cisco systems netflow services export version 9. [S.l.], 2004.

[22] SINGH, V. K.; DUTTA, K. Dynamic price prediction for amazon spot instances. In:IEEE. 2015 48th Hawaii International Conference on System Sciences. [S.l.], 2015. p.1513–1520.

[23] Pisa, P.S.; Costa, B.; Gonçalves, J.A.; Varela de Medeiros, D.S.; Mattos, D.M.F. A Private Strategy for Workload Forecasting on Large-Scale Wireless Networks. Information 2021, 12, 488. "https://doi.org/10.3390/info12120488"

[24] SCULLEY, D. Web-scale k-means clustering. In: ACM. Proceedings of the 19th international conference on World wide web. [S.l.], 2010. p. 1177–1178.

[25] BHOLOWALIA, P.; KUMAR, A. Ebk-means: A clustering technique based on elbow method and k-means in wsn. International Journal of Computer Applications, Citeseer,v. 105, n. 9, 2014.

[26] SYAKUR, M.; KHOTIMAH, B.; ROCHMAN, E.; SATOTO, B. Integration k-meansclustering method and elbow method for identification of the best customer profile cluster. In: IOP PUBLISHING. IOP Conference Series: Materials Science and Engineering.[S.l.], 2018. v. 336, n. 1, p. 012017.

[27] REIS, L.H.A., MAGALHÃES, L.C.S., DE MEDEIROS, D.S.V. et al. An Unsupervised Approach to Infer Quality of Service for Large-Scale Wireless Networking. J NetwSyst Manage 28, 1228–1247 (2020). "https://doi.org/10.1007/s10922-020-09530-3"

[28] XU, X. From cloud computing to cloud manufacturing. Robotics and computerintegrated manufacturing, Elsevier, v. 28, n. 1, p. 75–86, 2012.

[29] AL-ROOMI, M.; AL-EBRAHIM, S.; BUQRAIS, S.; AHMAD, I. Cloud computingpricing models: a survey. International Journal of Grid and Distributed Computing, Citeseer, v. 6, n. 5, p. 93–106, 2013.

[30] DUTTA, P.; DUTTA, P. Comparative study of cloud services offered by Amazon, Microsoft & Google. International Journal of Trend in Scientific Research and Development, v. 3, n. 3, p. 981–985, 2019.

[31] AU-YEUNG, B.; CHU, D.; ENFANTE, M.; LOGAN, G.; SAELEE, K. Industry analysis:Cloud computing. MIS Majors, v. 48, 2017.

[32] ORBAN, S. Ahead in the Cloud: Best Practices for Navigating the Future of Enterprise IT. [S.l.]: CreateSpace Independent Publishing Platform, 2017.

[33] PAHL, C.; JAMSHIDI, P. Microservices: A systematic mapping study. In: CLOSER(1). [S.l.: s.n.], 2016. p. 137–146.

[34] Grant Dasher, Ines Envid e Brad Calder. 2022. Architectures for Protecting Cloud Data Planes. In arXiv, 2201.13010.

[35] PISA, Pedro; MATTOS, Diogo. Autenticação Única nos Ambientes em Nuvem da Amazon Web Services com Integração de Usuários do Google Suite. In: WORKSHOP DE GESTÃO DE IDENTIDADES DIGITAIS, 9, 2019, São Paulo. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2019. p. 36-47. DOI: https://doi.org/10.5753/wgid.2019.14030.

[36] MARINA, Baby; MEMON, Irfana; ALVI, Fizza Abbas; NABI, Mairaj; RAJPER, Adnan Manzor; RAJPUT, Ubaidullah. A Study Towards Exploring Access Control Mechanisms and its Limitations in Cloud Computing. VAWKUM Transactions on Computer Sciences, [S. l.], v. 11, n. 1, p. 229–244, 2023. DOI: 10.21015/vtcs.v11i1.1473. Disponível em: https://vfast.org/journals/index.php/VTCS/article/view/1473.

[37] Ilia Shevrin e Oded Margalit. 2023. Detecting Multi-Step IAM Attacks in AWS Environments via Model Checking. In Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23). USENIX Association. Anaheim, CA, 6025-6042.

[38] Kumar, Uppala Vijay and REDDY, Dr. E.MADHUSUDHANA, Preventing Unauthorized Users from Accessing Cloud Data (May 15, 2023). Available at SSRN: https://ssrn.com/abstract=4448543 or "http://dx.doi.org/10.2139/ssrn.4448543"

[39] P. Gill, W. Dietl and M. Tripunitara, "Least-Privilege Calls to Amazon Web Services," in IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 3, pp. 2085-2096, 1 May-June 2023, doi: 10.1109/TDSC.2022.3171740.

[40] Xueqiang Wang, Yuqiong Sun, Susanta Nanda e XiaoFeng Wang. 2023. Credit Karma: Understanding Security Implications of Exposed Cloud Services through Automated Capability Inference. In Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23). USENIX Association. Anaheim, CA, 6007-6024.

[41] SCHNEIDER, F. B. Least privilege and more [computer security]. IEEE Security &Privacy, IEEE, v. 1, n. 5, p. 55–59, 2003.

[42] Miloslava Plachkinova & Kenneth Knapp (2023) Least Privilege across People, Process, and Technology: Endpoint Security Framework, Journal of Computer Information Systems, 63:5, 1153-1165, DOI: 10.1080/08874417.2022.2128937

[43] GALANTE, V. Practical role-based access control. Information Security Journal: AGlobal Perspective, Taylor & Francis, v. 18, n. 2, p. 64–73, 2009.

[44] LI, W.; WAN, H.; REN, X.; LI, S. A refined rbac model for cloud computing. In:IEEE. 2012 IEEE/ACIS 11th International Conference on Computer and Information Science. [S.l.], 2012. p. 43–48.

[45] SANDERS, M.; YUE, C. Automated least privileges in cloud-based web services. In: Proceedings of the fifth ACM/IEEE Workshop on Hot Topics in Web Systems andTechnologies. [S.l.: s.n.], 2017. p. 1–6.

[46] SANDERS, M. W.; YUE, C. Minimizing privilege assignment errors in cloud services. In: Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy. [S.l.: s.n.], 2018. p. 2–12.

[47] SANDERS, M. W.; YUE, C. Mining least privilege attribute based access controlpolicies. In: Proceedings of the 35th Annual Computer Security Applications Conference. [S.l.: s.n.], 2019. p. 404–416.

[48] HAN, J.; PEI, J.; YIN, Y.; MAO, R. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data mining and knowledge discovery, Springer, v. 8, n. 1, p. 53–87, 2004.

[49] NARAYANAN, H. A. J.; GÜNES, M. H. Ensuring access control in cloud provisioned healthcare systems. In: IEEE. 2011 IEEE Consumer Communications and Networking Conference (CCNC). [S.l.], 2011. p. 247–251.

[50] Boomija, M.D., Raja, S.V.K. Securing medical data by role-based user policy with partially homomorphic encryption in AWS cloud. Soft Comput 27, 559–568 (2023). "https://doi.org/10.1007/s00500-022-06950-y"

[51] Y. Hu, W. Wang, S. Khurshid, K. L. McMillan and M. Tiwari, "Fixing Privilege Escalations in Cloud Access Control with MaxSAT and Graph Neural Networks," 2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE), Luxembourg, Luxembourg, 2023, pp. 104-115, doi: 10.1109/ASE56229.2023.00167.

[52] Xiaomeng Lei and Mahesh Tripunitara. 2023. The Hardness of Learning Access Control Policies. In Proceedings of the 28th ACM Symposium on Access Control Models and Technologies (SACMAT '23). Association for Computing Machinery, New York, NY, USA, 133–144. "https://doi.org/10.1145/3589608.3593840"

[53] de Oliveira, N.R.; Pisa, P.S.; Lopez, M.A.; de Medeiros, D.S.V.; Mattos, D.M.F. Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges. Information 2021, 12, 38. https://doi.org/10.3390/info12010038

[54] B. B. A. da Costa and P. S. Pisa, & quot;Cloud Strategies for Image Recognition," 2020 4th. Conference on Cloud and Internet of Things (CIoT), Niteroi, Brazil, 2020, pp. 57-58. doi: 10.1109/CIoT50422.2020.9244200.