Estratégias de Redução de Custos em Nuvem sob a Perspectiva do Usuário: Um Mapeamento Sistemático

Elisa de Fátima Andrade Soares
efas@cin.ufpe.br
Universidade Federal de Pernambuco
Recife, PE

David Junio Mota Cavalcanti
djmc@cin.ufpe.br
Universidade Federal de Pernambuco
Recife, PE

Ioram Schechtman Sette
iss@cin.ufpe.br
Universidade Federal de Pernambuco
Recife, PE

Carlos André Guimarães Ferraz
cagf@cin.ufpe.br
Universidade Federal de Pernambuco
Recife, PE

ABSTRACT

Cloud-based systems operate under a pay-per-use pricing model, which brings challenges related to budget cost management. Cloud financial management, anchored in FinOps principles, balances performance with budget control. Existing cloud cost management solutions have explored different approaches, such as resource requirement estimation, historical data analysis, and automated scaling decisions, which are already widely adopted. However, existing studies still lack a comprehensive view of cost optimization strategies, which vary between user and provider perspectives, and there is no way to classify or choose the most appropriate approach for every scenario. In this context, this paper proposes a systematic mapping study covering cloud cost reduction strategies from the user's perspective. In particular, a complete methodology approach was defined by formulating detailed research questions, using carefully crafted search strings in scientific databases, and establishing organized by and structured by (processing, storage, network), strategy, type strategy, service type, cloud type and resource. Finally, emerging trends, research gaps, and practical implications for industry and academia are discussed.

KEYWORDS

cloud cost optimization, cloud computing, finOps, systematic mapping study

1 INTRODUÇÃO

A computação em nuvem transformou como as organizações constroem sistemas de software e gerenciam sua infraestrutura de TI. Ela permite migrar os sistemas para serviços online que oferecem disponibilidade, desempenho, escalabilidade e adaptabilidade [30, 100]. No entanto, por operarem sob um modelo de pagamento por uso, esses sistemas enfrentam desafios relacionados ao gerenciamento de custos [16, 30].

FinOps (*Financial Operations*) é a prática operacional que combina finanças e operações para otimizar os custos de nuvem dos sistemas. Ela busca promover o equilíbrio entre desempenho e controle orçamentário. Estratégias como o uso de instâncias *spot*, *autoscaling*, políticas de desligamento de recursos ociosos e análise preditiva de custos já fazem parte das melhores práticas em diversos setores [30, 118]. Segundo o relatório Flexera 2025 [45], a otimização de custos permanece, pelo sétimo ano consecutivo, como a principal iniciativa nas estratégias de nuvem das organizações.

Neste cenário, é essencial adotar estratégias inteligentes para alocação e gerenciamento da infraestrutura em nuvem. A complexidade crescente desses ambientes e a alta demanda por recursos exigem soluções que auxiliem na escolha otimizada de configurações e serviços. Por exemplo, ferramentas de recomendação baseadas em aprendizado de máquina são usadas para sugerir configurações adequadas a diferentes cargas de trabalho, considerando desempenho, confiabilidade e custo para sistemas [15, 92].

Mesmo com esses avanços, ambientes *multicloud* e híbridos, aliados à evolução de tecnologias como *serverless* e *edge computing*, aumentam a complexidade de gerenciamento e dificultam a previsibilidade de custos [106]. Além disso, a variedade de modelos de precificação oferecidos pelos provedores, tais como, sob demanda, reserva de instâncias e modelos baseados em leilão, requerem análise criteriosa para evitar desperdícios [16].

No entanto, a literatura carece de uma visão abrangente das estratégias de otimização de custos. Duas lacunas principais se destacam: (1) a ausência de um mapeamento sistemático que acompanhe a evolução das abordagens na última década; e (2) a escassez de classificações que integrem técnicas, ferramentas, tipos de serviços de nuvem, tipos de nuvem e contextos de adoção prática [16, 76].

Neste contexto, este artigo propõe preencher essas lacunas por meio de um mapeamento sistemático da literatura entre 2018 e 2024, com foco na categorização e análise de abordagens para redução de custos em ambientes de nuvem. Em particular, o estudo foca nas estratégias de redução de custos em nuvem sob a perspectiva do usuário e estabelece uma classificação evolucionária, organizada por geração tecnológica e estruturada por dimensões como tipo de recurso (processamento, armazenamento, rede), modelo de serviço e tipo de nuvem. Também são discutidas tendências emergentes, lacunas de pesquisa e implicações práticas para a indústria e a academia.

Finalmente, ao estruturar o conhecimento acumulado e revelar padrões evolutivos, este trabalho contribui como referencial teórico e prático para pesquisadores, arquitetos de soluções e gestores financeiros que atuam no ecossistema da computação em nuvem.

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

A estrutura deste artigo é composta por 7 seções. A Seção 2 introduz os conceitos básicos de otimização de custos em nuvem e descreve brevemente os tipos de nuvens. Em seguida, a Seção 3 apresenta detalhes da metodologia de Mapeamento Sistemático utilizada. A Seção 4 apresenta os resultados obtidos a partir do Mapeamento Sistemático. A Seção **??** identifica os riscos que podem comprometer a validade dos resultados da revisão. A Seção 6 aborda as lacunas existentes nas pesquisas e as perspectivas futuras em relação aos achados do estudo. Por fim, a Seção 7 apresenta as considerações finais, onde se realiza um sumário deste trabalho.

2 CONCEITOS BÁSICOS

Antes de apresentar a metodologia do mapeamento sistemático, é necessário introduzir os conceitos fundamentais sobre redução de custos em nuvem, além dos principais tipos de implantação e tipos de serviços de nuvem.

2.1 Estratégias de Redução de Custos

As estratégias de redução de custos em nuvem envolvem práticas que visam reduzir as despesas com infraestrutura, serviços e operações, mantendo o desempenho e disponibilidade. As estratégias são organizadas nas categorias de **processamento**, **armazenamento** e **rede**.

- **Estratégias de Processamento:** conjunto de técnicas para alocação, gerenciamento e otimização de recursos computacionais na nuvem, visando atender às demandas de desempenho ao menor custo possível. Essas técnicas incluem medidas para reduzir custos relacionados a instâncias de máquinas virtuais, contêineres e computação serverless [42, 43].
- **Estratégias de Armazenamento:** inclui técnicas para gerenciar o armazenamento, como compressão, deduplicação, migração para classes de armazenamento mais econômicas e políticas de ciclo de vida de dados [44].
- **Estratégias de Rede:** inclui técnicas para redução de custos relacionados à transferência de dados e comunicação, com minimização de tráfego, uso de redes de entrega de conteúdo (CDNs), balanceamento de carga e configuração de *Quality of Service* (QoS)[43].

2.2 Tipos de Nuvem

Nesta subseção, são discutidos os principais tipos de nuvem reconhecidos tanto na literatura quanto, na prática de mercado: nuvem privada, nuvem pública, nuvem híbrida e multicloud. Cada modelo é apresentado com sua respectiva definição e exemplo de aplicação. Conforme indicado por [82], a classificação dos tipos de nuvem constitui um elemento fundamental para a compreensão das estratégias de otimização de custos em ambientes de computação em nuvem. Os diferentes tipos de nuvem apresentam características distintas em relação à forma de provisionamento de recursos, ao grau de controle administrativo, à localização da infraestrutura e ao nível de compartilhamento entre usuários [81]. Esses aspectos influenciam diretamente fatores como elasticidade, desempenho, segurança, conformidade regulatória e, sobretudo, a estrutura de custos, elementos centrais para decisões de arquitetura e gestão financeira no contexto de *FinOps* [17, 118].

- **Nuvem Privada (Private Cloud)**: infraestrutura de nuvem dedicada exclusivamente a uma organização, podendo ser mantida localmente ou hospedada por terceiros [81]. *Exemplo*: um hospital que usa uma nuvem privada para gerenciar registros médicos sensíveis. *Diferencial*: controle total sobre os recursos e alto nível de segurança e privacidade.
- **Nuvem Pública (Public Cloud)**: serviços de nuvem disponibilizados por provedores terceirizados (como AWS, Google Cloud, Azure) para uso compartilhado entre múltiplos clientes, geralmente em um modelo pay-per-use [118]. *Exemplo*: uma startup usa o Google Cloud para implantar seu aplicativo sem precisar de infraestrutura própria. *Diferencial*: infraestrutura compartilhada, fácil escalabilidade e custo acessível.
- **Multicloud**: uso de múltiplos provedores de nuvem pública ou privada de forma independente, sem necessariamente haver integração entre eles [118]. *Exemplo*: uma empresa utiliza a AWS para armazenamento de dados, o Google Cloud para aprendizado de máquina e o Azure para serviços de aplicativos. *Diferencial*: não há interoperabilidade entre as nuvens, mas os serviços são utilizados conforme a necessidade.
- **Nuvem Híbrida (Hybrid Cloud)**: é um tipo específico de multicloud, caracterizada pela combinação de uma nuvem privada (local ou hospedada) com uma ou mais nuvens públicas, permitindo que os dados e aplicativos sejam compartilhados entre elas [81]. *Exemplo*: uma organização armazena dados sensíveis em sua nuvem privada local, mas utiliza a nuvem pública da Azure para escalar recursos em períodos de alta demanda. *Diferencial*: definida pela interoperabilidade e pela integração entre as nuvens pública e privada.

2.3 Tipos de Serviços de Nuvem

Os tipos de serviço em computação em nuvem constituem diferentes níveis de abstração e divisão de responsabilidade entre provedores e consumidores, exercendo influência direta sobre aspectos como flexibilidade, escalabilidade, custo e complexidade de gestão. A compreensão clara dessas modalidades é crucial para a identificação de oportunidades de otimização de custos, pois cada modelo impõe restrições e oferece vantagens específicas para distintos perfis de uso [81]. Em seguida, são descritos os três modelos de serviço amplamente reconhecidos pelo NIST [82]: *Software as a Service* Software como Serviço (SaaS), *Platform as a Service* Plataforma como Serviço (PaaS) e *Infrastructure as a Service* Infraestrutura como Serviço (IaaS), destacando suas características, responsabilidades e implicações para estratégias de *FinOps*.

- **Software como Serviço (Software-as-a-Service ou SaaS)**: modelo de entrega de software em que o provedor é responsável por hospedar, operar e administrar toda a aplicação em uma infraestrutura de nuvem multi-inquilino. O acesso é disponibilizado aos usuários finais através de interfaces *thin client*, tais como navegadores web, ou por meio de APIs. Nesse modelo, o consumidor não possui gestão ou controle sobre a infraestrutura subjacente, que inclui rede, servidores, sistema operacional, armazenamento ou os próprios componentes da aplicação, tendo acesso estritamente a funcionalidades limitadas, como a administração de contas de usuários,

configuração de preferências e manipulação dos dados de entrada e saída.

- **Plataforma como Serviço (Platform-as-a-Service ou PaaS)**: modelo de serviço em nuvem que oferece ao usuário uma plataforma abrangente, englobando infraestrutura, sistemas operacionais, ambientes de desenvolvimento e ferramentas para o desenvolvimento, teste, implantação e gerenciamento de aplicações. O usuário pode criar ou executar suas próprias aplicações, utilizando linguagens de programação, bibliotecas, serviços e ferramentas com suporte do provedor. A gestão da infraestrutura subjacente, como servidores, armazenamento, rede e sistemas operacionais, é de responsabilidade do provedor, permitindo que o consumidor se concentre exclusivamente na lógica e no ciclo de vida da aplicação [125].
- **Infraestrutura como Serviço (Infrastructure-as-a-Service ou IaaS)**: capacidade fornecida ao consumidor para prover processamento, armazenamento, rede e outros recursos básicos de computação, a partir dos quais é possível implantar e usar qualquer software, incluindo sistemas operacionais e aplicações. As vantagens incluem escalabilidade e inicialização rápida, enquanto os riscos envolvem modelo de preços, potencial bloqueio, segurança e privacidade.

2.4 Recursos em Nuvem

A seguir, apresentam-se as definições dos conceitos fundamentais associados aos tipos de serviços em nuvem, estabelecendo sua relação com as estratégias de otimização de custos:

- **Máquinas Virtuais (Virtual Machines)**: instâncias computacionais simuladas que operam em hardware físico compartilhado por meio da virtualização [81]. No contexto da nuvem, as VMs são unidades fundamentais de computação em ambientes IaaS, com recursos como CPU, memória e disco configuráveis. A otimização de custos pode ser feita por meio de práticas como redimensionamento (*resizing*), desligamento de VMs ociosas e uso de instâncias *spot* ou reservadas [17].
- **Armazenamento (Storage)**: serviços que abrangem armazenamento em bloco, armazenamento em arquivo e armazenamento em objeto. A otimização de custos nesse contexto envolve técnicas como compressão, deduplicação, migração entre classes de armazenamento (quente/frio), seleção estratégica de regiões e utilização de armazenamento em múltiplas nuvens ou na borda (*edge storage*) [76].
- **Serverless**: modelo de computação baseado em eventos, no qual os usuários não precisam gerenciar servidores. Os recursos são alocados dinamicamente, e o faturamento é baseado na execução de funções [50, 51, 119]. Pode ser altamente eficiente para cargas intermitentes, reduzindo custos com infraestrutura ociosa. Contudo, exige atenção à granularidade das funções e ao tempo de execução para evitar sobrecustos.
- **Contêiner**: encapsulam aplicações e suas dependências em um ambiente leve e portável [81]. São mais eficientes do que VMs em termos de sobrecarga, e seu uso pode ser otimizado com escalonamento automático, orquestração (por exemplo, Kubernetes) e agrupamento inteligente de *workloads*, reduzindo a subutilização de recursos [17].
- **Banco de Dados (Database)**: incluem serviços gerenciados como Amazon RDS, Azure SQL Database, entre outros. A otimização de custos envolve a escolha adequada do tipo de banco (relacional vs. NoSQL), uso de instâncias sob demanda ou reservadas, desligamento automático em períodos de inatividade e dimensionamento conforme a demanda [76, 118]. Além disso, replicações e backups devem ser planejados com eficiência para evitar custos excessivos.

3 METODOLOGIA

A metodologia de Mapeamento Sistemático utilizada na condução desta pesquisa, foi estruturada com base nas diretrizes estabelecidas por [62], garantindo um processo rigoroso e replicável para a identificação e análise de estratégias de otimização de custos em computação em nuvem.

O mapeamento foi dividido em três fases principais: fase de planejamento, fase da condução e fase dos resultados, conforme mostrado na Figura 1.

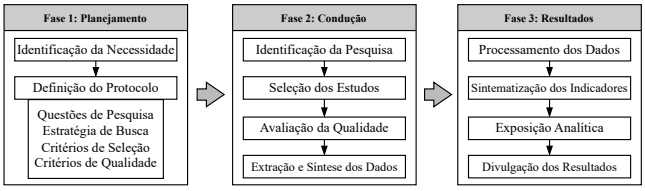

Figure 1: Fases do Mapeamento Sistemático

É importante destacar que esse mapeamento é predominantemente qualitativo, buscando compreender o tema por meio da extração, análise, classificação e interpretação das informações [25]. As subseções a seguir apresentam as fases do mapeamento sistemático em detalhes, incluindo todas as etapas e critérios estabelecidos.

3.1 Fase 1: Planejamento do Mapeamento Sistemático

A primeira fase corresponde ao planejamento, no qual são definidos os objetivos da pesquisa e elaborado o protocolo do mapeamento sistemático, incluindo a formulação das questões de pesquisa, os critérios de inclusão e exclusão, e a estratégia de busca. Essa fase é composta por duas etapas principais: (1) identificação da necessidade do Mapeamento Sistemático e (2) elaboração do protocolo, no qual são definidos os critérios metodológicos.

Etapa 1: Identificação da Necessidade do Mapeamento Sistemático. Conforme dito anteriormente, a complexidade inerente aos ambientes de computação em nuvem, derivada da diversidade de serviço (IaaS, PaaS, SaaS), tipos de nuvem (pública, privada, híbrida e multicloud) e estratégias de redução de custos, impõe desafios significativos à gestão eficiente de custos.

Deochake [30], apresenta algumas estratégias e técnicas de redução de custos de forma prática, acompanhadas de estudos de caso.

No entanto, ele adota uma abordagem narrativa que não segue os princípios formais de mapeamento sistemático. Por exemplo, ele não agrupa dimensões como tipos de serviços, nuvens e recursos limitando assim a generalização e a comparabilidade dos resultados.

Em contraste, o presente trabalho adota uma abordagem formal e reprodutível, reunindo estudos do período de 2018 a 2024, com foco na classificação das estratégias de redução de custos sob a perspectiva do usuário, além de identificar lacunas, tendências e implicações práticas, complementando e ampliando o trabalho do Deochake [30].

Etapa 2: Definição do Protocolo. Esta é uma das etapas mais importantes para o mapeamento sistemático, visto que o protocolo deve ser estabelecido antes mesmo de se iniciar a condução do mapeamento sistemático [40]. A Figura 2 apresenta as etapas do mapeamento sistemático.

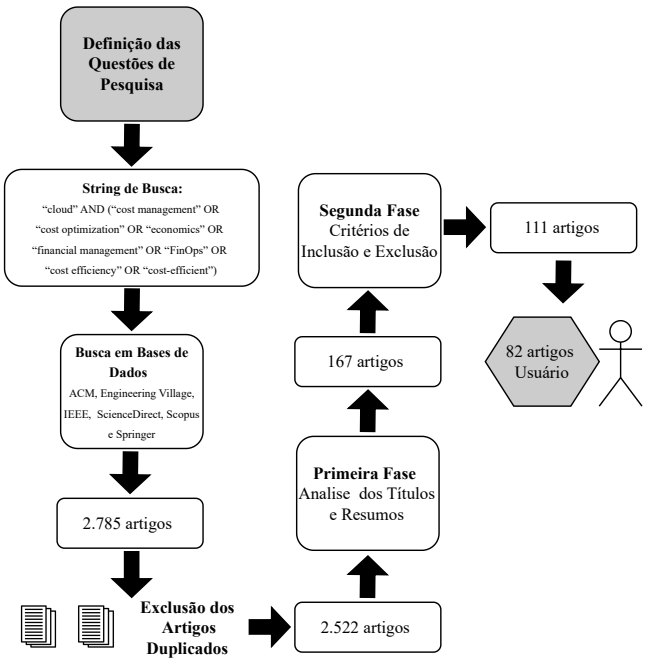

Figure 2: Processo de Seleção dos Estudos

Como pode ser observado, o protocolo define as questões de pesquisa, as *strings* de busca, as bases de dados e os critérios de inclusão, exclusão e qualidade, que auxiliam na seleção dos estudos relevantes e no descarte dos irrelevantes. Com o protocolo estabelecido, inicia-se a fase 2, dedicada à condução do mapeamento.

3.2 Fase 2: Condução do Mapeamento Sistemático

Nesta fase são realizadas as ações definidas no protocolo (Ver Figura 2). Conforme Figura 1, nesta fase são realizados os seguintes etapas: identificar da pesquisa, seleção dos estudos, avaliação da qualidade dos estudos, e extração e síntese dos dados extraídos.

O processo de execução do mapeamento sistemático teve início em 09 de janeiro de 2023 e se estendeu até 19 de maio de 2025, data em que foi finalizada a calibração das strings de busca e concluída a etapa de extração dos dados. Todo o protocolo seguiu as diretrizes consolidadas para estudos secundários na Engenharia de Software, garantindo rastreabilidade e reprodutibilidade das etapas.

Ressalta-se que não foi utilizada nenhuma ferramenta específica para gerenciamento de mapeamento sistemático ou revisão sistemática, como o Parsifal. Em vez disso, optou-se por utilizar Google Planilhas para registrar, organizar e consolidar as informações extraídas dos estudos selecionados, dada a flexibilidade da ferramenta para controle manual e visualização estruturada dos dados.

Etapa 1: Identificação da Pesquisa. Esta etapa, denominada indentificação da pesquisa, descreve o processo de elaboração das questões de pesquisa, fundamentais para direcionar a identificação e investigação. Kitcheham e Charters [62], recomendam o uso da estrutura PICOC (*Population*, *Intervention*, *Context*, *Outcomes* e *Comparison*). No entanto, Petticrew [99] apontam que, para estudos amplos e exploratórios, como mapeamentos sistemáticos, somente os elementos *população* e *intervenção* devem ser considerados. Neste estudo, *população* refere-se à redução de custos em computação em nuvem, enquanto *intervenção* diz respeito às estratégias para redução de custos.

Assim, o objetivo é compreender e avaliar as principais estratégias para reduzir custos em computação em nuvem, identificando indicadores de desempenho e explorando abordagens emergentes ainda pouco investigadas. Com base nisso, o mapeamento responde às seguintes questões:

Q1. Quais são as principais estratégias adotadas para a redução de custos em ambientes de computação em nuvem?
Q2. Como as estratégias de redução de custos em computação em nuvem se aplicam aos diferentes tipos de nuvem (privada, pública, híbrida e multicloud)?
Q3. Em quais tipos de serviços em nuvem (IaaS, SaaS e PaaS) são mais frequentemente implementadas essas estratégias de redução de custos?
Q4. Quais são os recursos de computação em nuvem (processamento, armazenamento e rede) são comumente utilizados nas estratégias de redução de custos em ambientes de nuvem?

Etapa 2: Seleção dos Estudos. Nesta etapa, os estudos são selecionados por meio de *strings* de buscas exaustivas, criadas a partir da combinação de palavras-chave e seus sinônimos. Esse método garante que todos os termos relevantes sejam considerados, evitando a perda de artigos importantes.

De acordo com Spanos e Angelis [110], a definição das strings segue um processo iterativo. Ele se inicia com buscas usando palavras-chave de artigos que abordam o contexto investigado e finaliza quando um conjunto inicial de artigos já conhecidos é encontrado pela busca efetuada. A estratégia adotada foi a busca automática em bases de dados bem conhecidas, tas como ACM, Engineering Village, IEEE, ScienceDirect, Scopus e Springer.

É importante mencionar que a construção das strings de busca foi realizada a partir dos seguintes processos:

(1) Investigação dos artigos relevantes no contexto de estratégias de redução de custos em nuvem e da extração das suas palavras-chave;
(2) As perguntas de pesquisa foram formuladas conforme a estrutura PICOC, considerando a *população* e *intervenção*;
(3) Identificaram-se os sinônimos dos termos principais;
(4) Utilizaram-se os conectores booleanos (AND e OR) para realizar a interconexão entre os termos;
(5) Realizou-se a verificação das strings de busca formuladas, as quais foram adaptadas e executadas nas fontes de dados consideradas nesta pesquisa.

Conforme a recomendação de Dyba e Dingsoyr [37], foram utilizados termos de busca abrangentes para maximizar a inclusão de estudos. A *string* de busca geral formulada foi:

"cloud" AND ("cost management" OR "cost optimization" OR "economics" OR "financial management" OR "FinOps" OR "cost efficiency" OR "cost-efficient")

Para refinar os resultados obtidos, a string geral foi adaptada para cada base de dados (ACM, Engineering Village, IEEE, ScienceDirect, Scopus e Springer) e foram definidos critérios de inclusão (CI) e exclusão (CE).

Os critérios de inclusão foram empregados para refinar a lista de artigos encontrados, sendo mantidos somente aqueles que serão considerados pertinentes à condução da pesquisa. Nesse contexto, limitações como o escopo específico, a língua e o período de análise figuram entre os fatores a serem considerados. Em relação aos critérios de exclusão, eles se aplicam aos artigos a serem descartados. Os critérios estabelecidos para o mapeamento sistemático proposto foram os seguintes: **Critérios de Inclusão (CI)**

CI1. Artigos que abordam estratégias de redução de custos em computação em nuvem;
CI2. Artigos publicados em inglês;
CI3. Artigos publicados entre 2018 e 2024.

Critérios de Exclusão (CE)

CE1. Artigos que não estão alinhadas ao escopo das questões de pesquisa definidas;
CE2. Artigos que abordam as estratégias de otimização de custos, seja sob a perspectiva do provedor;
CE3. Artigos de formato inadequado, como short papers, capítulos de livros, publicações de demonstração, estudos apresentados como resumos, apresentações em PowerPoint, teses e relatórios técnicos;
CE4. Artigos cujo acesso gratuito não esteja disponível nas bases de dados selecionadas.

Os critérios de inclusão definidos são os seguintes: o Critério de Inclusão CI1 estipula que os artigos selecionados devem utilizar estratégias visando a redução de custos em computação em nuvem, em consonância com o objetivo geral desta pesquisa. O critério CI2 determina a inclusão exclusiva de artigos escritos em inglês, dada sua importância no contexto acadêmico. Consequentemente, este mapeamento contempla artigos publicados a partir de 2018, em razão da necessidade identificada de expandir o corpus em análise. Portanto, artigos publicados no período de 2018 a 2024 são considerados para seleção.

3.2.1 Etapa 1: Identificação da Necessidade do Mapeamento Sistemático. Quanto aos critérios de exclusão, CE1 estabelece que serão removidos os estudos que não estejam alinhados ao escopo das questões de pesquisa definidas. O CE2 exclui trabalhos que abordem estratégias de otimização de custos sob a perspectiva do provedor de serviços em nuvem, uma vez que o foco desta pesquisa é a visão do consumidor. O CE3 determina a eliminação de estudos em formatos inadequados, como short papers, capítulos de livros, publicações de demonstração, resumos, apresentações em PowerPoint, teses e relatórios técnicos, de modo que a literatura cinzenta não seja considerada. Por fim, o CE4 exclui artigos cujo acesso gratuito não esteja disponível nas bases de dados selecionadas, a fim de facilitar a replicabilidade desta pesquisa.

A Figura 2, mostra os resultados da classificação da avaliação dos estudos selecionados. Ao executar as strings de busca, 2.785 artigos foram selecionados inicialmente. Desses, por estarem duplicados, 263 artigos foram eliminados. Em seguida, iniciou-se o processo de análise, que foi feito a partir da leitura do título e do resumo de cada um dos trabalhos. Logo após, realizou-se a segunda análise, considerando-se os critérios de inclusão e de exclusão estabelecidos. Por fim, procedeu-se à terceira análise, a qual pode ser descrita como a mais detalhada, visto que se realizou a leitura completa dos artigos, de modo a encontrar as evidências das questões de pesquisa levantadas. Essa etapa resultou em 112 artigos finais, sendo 82 na perspectiva do usuário, o foco deste artigo e os 29 artigos na perspectiva do provedor que foram excluídos. Por fim, após a seleção dos estudos relevantes, foi efetuada uma avaliação da qualidade dos artigos.

3.2.2 Etapa 3: Avaliação da Qualidade do Estudo. Para [63], a avaliação da qualidade dos artigos é importante, tendo em vista que esse procedimento analisa e verifica a alta qualidade do nível dos artigos incluídos em um mapeamento sistemático. Em conformidade com essa colocação, [60] afirmam, em seu estudo, que as avaliações da qualidade são consideradas um processo que determina a credibilidade dos estudos e uma maneira de atribuir notas ao conteúdo gerado. Visando que as questões avaliativas examinam as pesquisas com o intuito de minimizar o viés da pesquisa. Os Critérios de Avaliação (CA) para averiguar a qualidade dos estudos são definidos pelas seguintes perguntas:

CA1. O artigo aborda as **estratégias** utilizadas para reduzir os custos na nuvem de forma detalhada e prática?
CA2. O artigo apresenta **orientações** específicas ou soluções aplicáveis para a implementação das estratégias de redução de custos em nuvem?
CA3. O **método** adotado no estudo é descrito de maneira clara e detalhada?
CA4. Os **resultados** apresentados são analisados e interpretados de forma compreensível e lógica?
CA5. O **ambiente de teste** usado nos experimentos está claramente definido?

A avaliação dos artigos foi conduzida utilizando-se uma escala Likert de cinco pontos, amplamente reconhecida por sua eficácia na mensuração de percepções qualitativas [9, 24, 31]. De acordo com [9], esse formato proporciona discriminação adequada sem introduzir complexidade excessiva; [24] observa que essa abordagem diminui o viés de resposta central e [31] evidencia que há

um incremento na confiabilidade entre os avaliadores. A pontuação final foi calculada por meio da média aritmética de cinco critérios, classificando-se nas categorias: Excelente (4,5–5,0), Muito Bom (4,0–4,5), Bom (3,0–4,0), Regular (2,0–3,0) e Ruim (1,0–2,0).

3.2.3 Etapa 4: Extração e Síntese de Dados. Esta etapa consiste na análise detalhada das informações extraídas da amostra final dos artigos selecionados. Somente os conteúdos que respondem diretamente às questões de pesquisa estabelecidas são considerados relevantes. Para assegurar a organização dos dados e o controle de versões, foi desenvolvido no Google Planilhas, onde todas as informações foram registradas de forma estruturada. A seguir, são apresentados os campos utilizados para o mapeamento das informações:

(1) **Título da Publicação:** descrição sucinta e clara do tema principal abordado no trabalho.
(2) **Resumo da Publicação:** apresenta os principais objetivos, métodos utilizados e conclusões alcançadas no estudo.
(3) **Ano de Publicação:** indica o ano em que o trabalho foi publicado, para contextualizar sua relevância temporal.
(4) **Fonte de Busca:** identificação da base de dados, plataforma ou ferramenta empregada para localizar artigos, tais como ACM, IEEE, e Scopus.
(5) **Palavras-chave:** lista de termos ou expressões que representam os principais temas abordados na publicação.
(6) **Autor Principal:** nome do pesquisador ou pesquisadora que liderou o desenvolvimento do trabalho.
(7) **Autores:** relação completa dos autores que contribuíram para a realização da pesquisa.
(8) **Fonte da Publicação:** nome do periódico, conferência ou outro meio onde o trabalho foi publicado.
(9) **Tipo de Fonte de Publicação:** classificação do tipo de publicação, como artigo de conferência ou artigo de periódico.
(10) **Estratégia:** seria a estratégia de custo utilizada para obter a redução de custo em nuvem
(11) **Tipos de Estratégia:** refere-se às diversas abordagens estratégicas empregadas para a redução de custos na computação em nuvem, que incluem processamento, armazenamento e rede.
(12) **Tipo de Nuvem:** indica o tipo de ambiente de nuvem utilizado no estudo, como pública, privada, híbrida ou multicloud.
(13) **Modelo (IaaS, PaaS, SaaS):** Detalhamento do modelo de serviços em nuvem no qual a estratégia de redução foi aplicada.
(14) **Serviço (Máquinas Virtuais, Contêineres, Funções Serverless, entre outros):** especificação do serviço em nuvem que passou por redução de custos.

3.3 Fase 3: Resultados do Mapeamento Sistemático

A última fase envolve a documentação da pesquisa já concluída, devidamente organizada, com a apresentação dos resultados obtidos a partir da análise dos estudos selecionados. Os achados são sistematizados de modo a proporcionar uma visão abrangente do estado atual das estratégias de otimização de custos em ambientes de computação em nuvem. Foram identificadas e categorizadas várias abordagens utilizadas para a redução de custos associada ao uso de recursos em nuvem, abrangendo desde técnicas de alocação de recursos e dimensionamento dinâmico até a utilização de instâncias spot.

4 RESULTADOS DO MAPEAMENTO SISTEMÁTICO

Nesta seção, são expostos os resultados obtidos no Mapeamento Sistemático. A partir desses resultados, serão discutidas as principais estratégias já existentes para a redução de custos em ambientes de computação em nuvem.

4.1 Distribuição Temporal das Publicações

A análise referente à distribuição temporal das publicações no período de 2018 a 2024 apresentada na Figura 3 evidencia a evolução das investigações no âmbito da otimização de custos na computação em nuvem, tanto em termos de volume anual quanto nas fontes de indexação. Durante esse período, foram observados momentos de crescimento, picos na produção e fases de consolidação, acompanhando a introdução e o amadurecimento do FinOps (*Cloud Financial Operations*), que integra práticas técnicas e de gestão financeira para otimizar o consumo de recursos em nuvem.

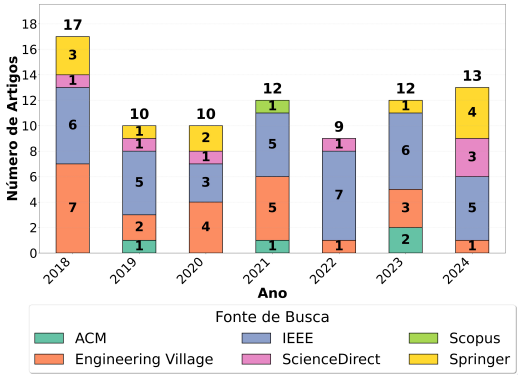

Figure 3: Distribuição por Bases de Dados

De 2018 a 2020, observou-se uma evolução inicial nas publicações sobre redução de custos em computação em nuvem, com oscilações anuais e mudanças no perfil das contribuições. Em 2018, foram coletados com 17 artigos, predominando estudos técnicos e aplicados, liderados pela Engineering Village e IEEE Xplore. Já em 2019, o volume caiu para 10 artigos com um foco maior na formalização de métodos e na inserção em eventos e periódicos de engenharia. Em 2020, foram coletadas 10 publicações, com maior equilíbrio entre IEEE Xplore e Engineering Village, indicando um momento de transição, marcado pela coexistência de abordagens práticas e estudos voltados à avaliação de desempenho e padronização.

Já em 2021, notou-se uma recuperação, totalizando 12 publicações. Esse ano coincidiu com uma disseminação mais abrangente do conceito de FinOps, promovido pela FinOps Foundation, que favoreceu a integração entre métricas técnicas e análise de custos para decisões estratégicas. O pico de especialização surgiu em 2022,

com a IEEE Xplore dominante e foco em métricas de alocação, modelos preditivos e governança. Em 2023, o volume de artigos cresceu para 12, possivelmente impulsionado pela chegada de novos temas, como *Green Cloud Computing* e o uso de inteligência artificial para previsão e gerenciamento de custos. Por fim, em 2024, manteve-se alto (13 artigos), com maior diversificação de fontes e síntese do conhecimento acumulado, evidenciando a maturidade das práticas de FinOps.

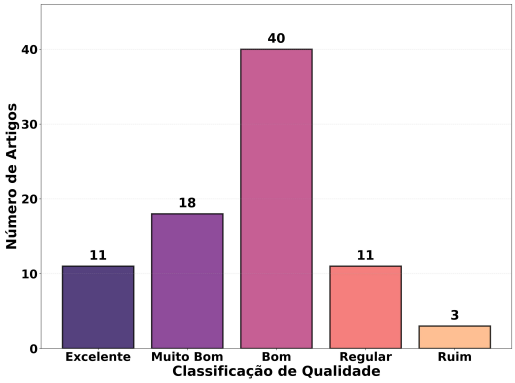

Figure 4: Avaliação de Qualidade

4.2 Avaliação da Qualidade dos Artigos

A avaliação da qualidade foi conduzida conforme os critérios previamente definidos na Seção 3. Dos 82 artigos avaliados ilustrados na Figura 4, verificou-se que a maioria apresentou qualidade considerada boa, correspondendo a 40 estudos (48,8%). Em seguida, destacaram-se as categorias muito boa, com 18 estudos (22,0%), e excelente, com 11 estudos (13,4%). As classificações menos favoráveis, regular e ruim, representaram, respectivamente, 11 (13,4%) e 3 (3,7%) trabalhos. Esses resultados evidenciam que, de forma geral, o conjunto analisado possui qualidade de satisfatória a elevada, com predominância de estudos bem estruturados e metodologicamente consistentes.

4.3 Análise das Evidências Mapeadas

A análise das evidências provenientes do mapeamento sistemático revelou um conjunto extenso de estratégias de otimização de custos aplicáveis a diversos recursos de computação em nuvem. As subsequentes seções apresentam o mapeamento dessas estratégias com o intuito de abordar as questões formuladas no estudo. Tais estudos podem ser classificadas conforme a própria estratégia, seu tipo, o tipo de serviço em nuvem e o recurso-alvo para a redução de custos, conforme ilustrado na Tabela 1 para cada artigo analisado.

Para fins de padronização e concisão, os elementos apresentados na tabela são representados por meio de **siglas** que correspondem às categorias utilizadas no processo de análise. Essa abordagem permite condensar informações complexas de maneira sistemática, favorecendo a visualização comparativa entre estudos com diferentes escopos e enfoques. Além disso, o uso de abreviações facilita a identificação de padrões recorrentes nas estratégias analisadas e contribui para uma leitura mais objetiva e estruturada dos resultados obtidos.

Table 1: Estratégias de Redução de Custos em Nuvem

Ref.	Estratégia	Tipo Estratégia	Serviço	Nuvem	Recurso
[1, 8, 14, 19, 21, 23, 39, 73, 101, 114, 117, 128]	AJ	PR	IaaS	Pública	VM
[4]	AJ	PR	IaaS	Híbrida	VM
[53, 104]	AR	PR	IaaS	Híbrida	VM
[22, 27, 32, 33, 36, 84, 109, 111, 127]	AR	PR	IaaS	Pública	VM
[28, 57]	AR	PR	IaaS	Multicloud	VM
[26]	AR	PR	IaaS	Multicloud	VM, ARM
[95]	AR	PR	IaaS/PaaS/SaaS	Pública	VM, BD
[71]	AR	PR	IaaS/PaaS/SaaS	Híbrida	VM, BD
[59]	AS	PR	IaaS	Multicloud	VM, ARM
[91, 102]	AS	PR	IaaS	Pública	VM
[47]	AS	PR	IaaS	Pública	VM, ARM
[54]	AS, CT	PR	PaaS	Pública	VM, CT
[116]	AS, CT	PR	IaaS	Pública	VM, CT
[2]	AS, BC	RD	IaaS	Pública	VM, BC
[100]	AS, SV, CT	PR	PaaS	Pública	VM
[107]	BC	RD	IaaS	Híbrida	VM, BC
[88]	OA	ARM	IaaS	Pública	ARM
[10, 18, 35, 49, 56, 66, 67, 78, 86, 93, 94]	IS	PR	IaaS	Pública	VM
[124]	RI	PR	IaaS	Pública	VM
[20]	RI, AR	PR	IaaS/SaaS	Pública	VM
[48]	RI, AR	PR	IaaS	Pública	VM
[70]	RI, AS	PR	IaaS	Híbrida	VM
[96]	IS	PR	IaaS	Híbrida	VM, CT
[3, 108, 112]	IS	PR	IaaS	Pública	VM
[87]	IS, AS	PR	IaaS	Pública	VM
[7]	IS, RI	PR	IaaS	Pública	VM
[75]	OA	ARM	PaaS	Pública	ARM
[5, 76, 79]	OA	ARM	IaaS	Pública	ARM
[34]	OA, AR	ARM	IaaS	Pública	VM, ARM
[12]	CT	PR	IaaS	Pública	VM, CT
[13]	CDN	RD	IaaS	Pública	VM
[46, 89, 90, 121–123]	RI	PR	IaaS	Pública	VM
[68]	RI	PR	IaaS	Multicloud	VM
[72, 74]	SV	PR	PaaS	Pública	VM
[38]	SV	PR	PaaS	Pública	VM, SV
[85]	SV	PR	PaaS	Pública	VM, ARM
[58]	TI	ARM	IaaS	Multicloud	ARM
[41]	TI	ARM	IaaS	Pública	ARM
[58]	TI, PCV	ARM	IaaS	Pública	ARM

Estratégia — AJ: Agendamento de Jobs; AR: Alocação de Recursos; AS: Auto Scaling; BC: Balanceador de Carga; OA: Otimização de Armazenamento; IS: Instâncias Spot; RI: Reserva de Instâncias; SV: Serverless; TI: Tiering Inteligente; CT: Contêineres; CDN: Rede de Distribuição de Conteúdo; PCV: Políticas de Ciclo de Vida.
Tipo Estratégia — PR: Processamento; ARM: Armazenamento; RD: Rede.
Recurso — VM: Máquina Virtual; ARM: Armazenamento; SV: Serverless; CT: Contêiner; BC: Balanceador de Carga; BD: Banco de Dados.

4.3.1 Principais estratégias para a redução de custos quanto ao tipo de recurso. Nesta subseção, as estratégias identificadas foram classificadas em três categorias principais: **processamento**, **armazenamento** e **rede**. Entre os 82 artigos analisados, 70 (aproximadamente 85,4%) concentraram-se em técnicas voltadas ao **processamento**, 9 (cerca de 11,0%) abordaram estratégias relacionadas ao **armazenamento** e 3 (aproximadamente 3,7%) trataram de otimizações na

rede. As estratégias identificadas e discutidas nos estudos revisados serão apresentadas a seguir, acompanhadas por descrições e exemplos de aplicação prática. Esta abordagem visou oferecer uma perspectiva sistemática e comparativa, auxiliando pesquisadores e profissionais na seleção das soluções mais adequadas para diferentes contextos de aplicação.

Estratégias de Processamento. As estratégias desta categoria focaram na otimização do uso de recursos computacionais, incluindo instâncias virtuais, contêineres e funções sob demanda. O objetivo central é alinhar a capacidade provisionada à demanda real, evitando ociosidade e reduzindo custos, sem comprometer o desempenho das aplicações. Entre os 82 artigos analisados, identificaram-se as seguintes frequências por tipo de estratégia: **Alocação de Recursos** (19 artigos), **Agendamento de *Jobs*** (13 artigos), **Instâncias *Spot*** (17 artigos), **Instâncias Reservadas** (12 artigos), ***Auto Scaling*** (11 artigos), **Otimização de Contêineres** (5 artigos) e ***Serverless*** (5 artigos). Ressalta-se que o termo *Jobs* foi mantido por uma questão estética e de padronização da tabela de resultados, referindo-se, na prática, ao agendamento de tarefas, e que "Contêineres" corresponde às estratégias de *otimização de contêineres*.

A predominância de estudos sobre **Alocação de Recursos** relacionou-se ao seu caráter essencial e aplicabilidade em diversos cenários de computação em nuvem. Como envolve o gerenciamento direto de CPU, memória, I/O e rede, essa abordagem é útil para reduzir custos em cargas de trabalho de diferentes portes e naturezas, além de poder, ser combinada com outras estratégias como *auto scaling* e instâncias *spot*, ampliando o impacto financeiro [16, 95, 113, 115, 118].

O **Agendamento de *Jobs*** apresentou presença intermediária, por depender de *workloads* com flexibilidade temporal para execução. Essa técnica é efetiva quando é possível programar tarefas para horários de menor custo ou baixa demanda, aproveitando janelas tarifárias reduzidas. A adoção tende a ser menor em aplicações que exigem processamento contínuo ou tempo de resposta imediato [4, 8, 92, 117, 128].

As **Instâncias *Spot*** apareceram com alta frequência devido ao seu elevado potencial de economia, já que oferecem descontos expressivos para *workloads* tolerantes a falhas. Sua aplicação é mais comum em cenários de processamento em lote, testes e ciência de dados. Apesar das vantagens, o risco de interrupção limita seu uso em serviços críticos, o que impede que essa estratégia seja predominante absoluta [18, 30, 49, 67, 86].

As **Instâncias Reservadas** foram relevantes pela previsibilidade e segurança que oferecem em cargas estáveis. Garantem economia em médio e longo prazo e são amplamente usadas por organizações com padrões de consumo regulares. Por outro lado, a rigidez contratual e a menor flexibilidade operacional explicam porque aparecem menos que as instâncias *spot* [7, 46, 118, 121, 124].

O ***Auto Scaling*** apresentou frequência equivalente às instâncias reservadas, por ser eficaz em ambientes com variação de demanda, ajustando automaticamente a capacidade provisionada. Apesar de ser um recurso nativo de muitos provedores, ele geralmente não é analisado isoladamente na literatura, sendo frequentemente incorporado a estratégias amplas de gerenciamento de recursos [16, 54, 70, 87].

A **Otimização de Contêineres** apresentou baixa frequência porque, embora reduza a sobrecarga de provisionamento em relação às VMs, o custo dos contêineres em muitos provedores ainda é indiretamente vinculado à infraestrutura subjacente. Isso faz com que seja percebida mais como uma técnica complementar do que como uma estratégia primária de redução de custos [12, 42, 116].

Por fim, o ***Serverless*** também apresentou baixa ocorrência, pois o modelo já incorpora cobrança proporcional ao uso e gerenciamento automático de escalabilidade pelo provedor. Dessa forma, há pouco espaço para intervenções diretas do usuário visando otimização, explicando seu uso restrito em estudos específicos [50, 51, 72, 74].

Estratégias de Armazenamento. As estratégias desta categoria têm como foco a otimização dos custos relacionados à conservação e ao acesso de dados em nuvem, explorando diferentes classes de serviço, políticas de retenção, técnicas de compressão e deduplicação. O objetivo é manter a disponibilidade e integridade das informações enquanto se reduz o gasto por meio do uso racional das camadas e regiões de armazenamento. Entre os 82 artigos analisados, identificaram-se as seguintes frequências: **Otimização de Armazenamento** (6 artigos), ***Tiering* Inteligente** (3 artigos) e **Políticas de Ciclo de Vida** (1 artigo).

A maior ocorrência de estudos sobre **Otimização de Armazenamento** está associada à sua abrangência e aplicabilidade em diferentes contextos, combinando técnicas como compressão, deduplicação e escolha de classes de armazenamento adequadas para reduzir custos sem comprometer o desempenho [30, 76].

O ***Tiering* Inteligente** apresenta ocorrência moderada entre as estratégias identificadas sendo empregado para transferir dados automaticamente entre classes de armazenamento conforme a frequência de acesso, otimizando o custo de retenção de grandes volumes [30, 76].

As **Políticas de Ciclo de Vida** foram menos frequentes, mas exercem papel relevante na automação da movimentação ou exclusão de dados obsoletos, reduzindo gastos com armazenamento prolongado [76].

Estratégias de Rede. As estratégias desta categoria têm como foco a otimização do uso de recursos de conectividade e distribuição de tráfego, incluindo balanceamento de carga, replicação de conteúdo e minimização de transferências de dados. O objetivo central é reduzir custos associados a egressos e infraestrutura de rede, mantendo a qualidade de serviço, baixa latência e alta disponibilidade. Entre os 82 artigos analisados, identificaram-se as seguintes frequências: **Balanceador de Carga** (2 artigos) e **CDN** (1 artigo). Ressalta-se que, embora sejam menos abordadas que as estratégias de processamento e armazenamento, elas desempenham papel relevante em arquiteturas distribuídas, muitas vezes sendo aplicadas em conjunto com outras práticas de otimização.

A relevância dos estudos sobre **Balanceador de Carga** justificam-se por sua importância na distribuição eficiente do tráfego, prevenindo a sobrecarga de instâncias e evitando o provisionamento excessivo. Essa técnica é amplamente adotada em ambientes com variação de demanda e serviços críticos, sendo fundamental para assegurar alta disponibilidade e escalabilidade [18, 30, 49].

Já a **CDN** também proporciona benefícios relevantes para aplicações com alto volume de conteúdo estático ou multimídia. Ao

replicar dados em pontos de presença próximos aos usuários, reduz-se a latência e o tráfego na origem, resultando em economia de custos de transferência e melhorando a experiência do usuário [16, 30].

4.3.2 Estratégias de Redução de Custos quanto ao Tipo de Nuvem. No mapeamento sistemático da literatura indicado na Tabela 1, observou-se que, dos artigos analisados, 72 (aproximadamente 84,7%) abordaram a implementação de estratégias de redução de custos em ambientes de **Nuvem Pública**. Tal predominância está em consonância com a ampla adoção deste tipo de nuvem por organizações de diversos tamanhos e setores, devido à sua escalabilidade, disponibilidade global e modelo de cobrança baseado em demanda. Contudo, essas vantagens também acarretam riscos significativos de custos excessivos, especialmente na ausência de governança adequada dos recursos contratados. Esta situação impulsiona o desenvolvimento e aplicação de estratégias avançadas de otimização, tais como a utilização de instâncias *spot* e reservadas, *autoscaling*, agendamento de tarefas e políticas de desligamento automático de recursos ociosos, frequentemente identificadas nos estudos revisados. A alta competitividade do mercado e a volatilidade da demanda por recursos em nuvem pública reforçam a necessidade dessas práticas, favorecendo sua recorrência na literatura especializada.

Em segundo lugar, foram identificados 7 artigos (aproximadamente 8,2%) que tratam de estratégias direcionadas para **Nuvem Híbrida**. Este tipo de nuvem integra recursos de nuvens públicas e privadas, proporcionando flexibilidade aprimorada para a alocação de cargas de trabalho e melhor atendimento aos requisitos de segurança e conformidade. No entanto, sua complexidade operacional, associada à necessidade de integração e orquestração entre diferentes ambientes, torna mais desafiadora a implementação de estratégias de otimização. Mesmo assim, os estudos encontrados sugerem que, quando bem planejada, a adoção da nuvem híbrida possibilita ganhos significativos ao direcionar cargas sensíveis para a infraestrutura privada e *workloads* elásticos para a nuvem pública, maximizando o custo-benefício em cada contexto.

A ***multicloud*** aparece em 6 artigos (cerca de 7,1%), sendo caracterizada pelo uso simultâneo de múltiplos provedores de nuvem para reduzir riscos de dependência de fornecedor e aproveitar oportunidades de arbitragem de preços. Embora esse tipo de nuvem ofereça potencial para otimização de custos, os estudos analisados evidenciaram que sua adoção ainda é limitada devido à elevada complexidade de gerenciamento, necessidade de ferramentas especializadas para monitoramento unificado e desafios de compatibilidade entre serviços. Em contrapartida, quando bem aplicada, a *multicloud* permite explorar ofertas promocionais e otimizar a alocação de cargas em função de variações de custo entre provedores.

Por fim, a investigação revelou a ausência de estudos dedicados exclusivamente a ambientes de nuvem privada. Apesar da identificação de estudos que abordam nuvens híbridas (as quais combinam nuvens privadas e públicas), não foi encontrado nenhum trabalho cujo foco recaia exclusivamente sobre nuvens privadas. A escassez de pesquisas específicas pode ser explicada não só pela menor flexibilidade na implementação de estratégias de otimização de custos, já que, neste modelo, os custos frequentemente se associam a investimentos fixos em infraestrutura própria ou a contratos de longo prazo, mas também pela metodologia deste mapeamento, que prioriza a visão do usuário em vez da do provedor. Em contextos de nuvem privada, a gestão e o controle dos recursos são responsabilidade integral da própria organização, sendo que as ações de otimização são comumente dirigidas à eficiência operacional e energética, em vez da redução direta a custos financeiros.

4.3.3 As Estratégias de Redução de Custos por Tipos Serviços em Nuvem. Neste mapeamento sistemático da literatura, conforme apresentado na tabela 1 constatou-se que dos 82 artigos analisados, 72 (aproximadamente 86,7%) tratam de estratégias de redução de custos implementadas na camada **IaaS (*Infrastructure as a Service*)**. Essa predominância se verifica porque, neste tipo de serviço, os usuários detêm controle direto sobre recursos essenciais, como máquinas virtuais, armazenamento e rede, tarifados com base no uso efetivo. Essa visibilidade financeira detalhada possibilita a implementação de técnicas avançadas de otimização, tais como *rightsizing*, *Auto Scaling*, utilização de Reserva de Instâncias ou *spot*, além de políticas de desativação de recursos ociosos. Dessa forma, a IaaS se configura como o ambiente com maior potencial para intervenções personalizadas destinadas à economia de custos.

Outro aspecto que contribui para esta predominância é o elevado nível de flexibilidade e autonomia proporcionado pela IaaS. Os usuários conseguem configurar e gerenciar diretamente sua infraestrutura, algo que não é viável nos serviços superiores, como PaaS e SaaS. Além disso, a IaaS frequentemente serve como ponto inicial para organizações que migram seus sistemas para a nuvem, particularmente quando enfrentam aplicações legadas e cargas de trabalho previsíveis. Este cenário favorece o desenvolvimento de um corpo consistente de estudos, ferramentas e boas práticas voltadas especificamente para a otimização de custos nesta camada. A abordagem FinOps, que vem gradualmente adquirindo relevância, demonstra-se altamente compatível com o contexto da IaaS, onde a disponibilidade de métricas detalhadas de consumo e desempenho facilita a implementação de ações contínuas de controle financeiro.

Embora essa predominância seja evidente, o mapeamento também revelou que 8 artigos (cerca de 9,6%) abordam estratégias direcionadas à camada **PaaS (*Platform as a Service*)**. Trata-se de uma modalidade que fornece ao usuário uma plataforma abrangente para o desenvolvimento, teste, implantação e gerenciamento de aplicações. O usuário tem à disposição linguagens de programação, bibliotecas e ferramentas providas pelo fornecedor, sem a necessidade de gerenciar a infraestrutura subjacente, como servidores, armazenamento e sistemas operacionais. Esse modelo promove uma otimização de custos por padrão, particularmente em aplicações modernas que utilizam microsserviços ou funções orientadas a eventos. Todavia, dado que o controle direto da infraestrutura é transferido para o provedor, as oportunidades de aplicar estratégias detalhadas de otimização acabam sendo mais limitadas em comparação à IaaS.

Em relação ao serviço **SaaS (*Software as a Service*)**, apenas 3 artigos (aproximadamente 3,6% do total) foram identificados. Neste modelo de entrega mais abstrato, a responsabilidade de hospedar, operar e gerenciar o aplicativo na íntegra recai integralmente sobre o provedor. O acesso ao *software* ocorre por meio de interfaces *web* (*thin clients*) ou APIs, e o usuário final não dispõe de qualquer controle sobre os componentes de infraestrutura ou sobre a própria

aplicação. Tal condição limita sobremaneira as possibilidades de intervenção direta para otimização de custos, restringindo-se a ajustes elementares, como configurações de contas e manipulação de dados. Em virtude deste cenário, observa-se uma escassez de estudos focados em estratégias de otimização aplicadas à camada SaaS na literatura analisada.

4.3.4 As Estratégias de Redução de Custos quanto Tipos de Recurso em Nuvem. Conforme apresentado na Tabela 1, constatou-se que, dos artigos analisados, 74 (aproximadamente 90,2%) abordam estratégias de redução de custos voltadas à otimização de **Máquinas Virtuais (VM)**. Essa predominância é justificada pelo fato de que as VMs constituem o principal elemento de consumo em ambientes de nuvem, especialmente na camada IaaS, sendo diretamente associadas ao custo de processamento, memória e uso de rede. Por serem recursos com cobrança baseada no tempo de uso e na capacidade provisionada, oferecem amplo espaço para aplicação de técnicas como *right-sizing*, *auto scaling*, agendamento de *jobs*, uso de instâncias *spot* ou reservadas, e políticas de desligamento automático de recursos ociosos. Essa alta representatividade reforça que a otimização de VMs é o ponto central nas estratégias de FinOps e gestão de custos em nuvem, já que pequenas variações na configuração ou no tempo de execução podem gerar economias significativas.

O segundo recurso mais frequente foi o **Armazenamento**, presente em 13 artigos (15,8%). Embora menos recorrente que as VMs, o armazenamento é um componente crítico e recorrente nas estratégias de custo, especialmente considerando que seu consumo tende a crescer de forma contínua ao longo do tempo. As técnicas identificadas incluem deduplicação, compressão, *tiering* inteligente, políticas de ciclo de vida e otimização do tráfego de leitura/gravação para reduzir custos operacionais e de transferência de dados. No entanto, o foco menor em relação às VMs sugere que, na perspectiva do usuário final, o armazenamento é tratado mais como um recurso de suporte do que como alvo primário de otimização.

O **Banco de Dados (BD)** foi citado em apenas 2 artigos (2,4%), aparecendo principalmente em cenários PaaS e *multicloud*, onde ajustes de configuração, escolha de instâncias otimizadas e uso de serviços gerenciados são explorados para reduzir custos. Apesar de sua relevância operacional, o baixo número de estudos dedicados a BD reforça a percepção de que a otimização de banco de dados é muitas vezes incorporada como parte de estratégias mais amplas de aplicação, e não como foco exclusivo.

O uso de **contêineres** foi identificado em 5 artigos (6,1%), associado principalmente a estratégias de escalabilidade elástica e migração de *workloads*. Embora contêineres tragam vantagens como portabilidade e menor sobrecarga de provisionamento em relação a VMs, sua baixa presença na amostra pode estar ligada ao fato de que, no modelo de cobrança da maioria dos provedores, o custo do contêiner continua indiretamente atrelado à VM subjacente, diluindo sua identificação como recurso independente de otimização.

Por fim, o **Balanceador de Carga (BC)** foi encontrado em 2 artigos (2,4%), ligado a estratégias de otimização de rede, distribuição inteligente de tráfego e desligamento de instâncias ociosas. Já o *serverless* foi identificado como recurso em apenas 3 artigos (3,7%), o que pode ser explicado pelo fato de que, apesar de ser uma abordagem com cobrança estritamente proporcional ao uso, o foco de otimização nesse modelo é geralmente menor, já que o provedor gerencia automaticamente a escalabilidade e a alocação de recursos.

5 AMEAÇAS À VALIDADE

O mapeamento sistemático apresenta **ameaças à validade** que podem influenciar tanto a abrangência dos estudos incluídos quanto a confiabilidade das conclusões obtidas. A primeira ameaça está associada à **seleção dos estudos primários**, potencialmente sujeita a vieses de idioma e de base de dados, uma vez que a busca foi restrita a artigos publicados em inglês. Embora a *string* de busca tenha sido construída com base em trabalhos anteriores e revisada por especialistas, ela pode não ter contemplado todos os termos relevantes. Além disso, a aplicação da busca apenas em títulos, resumos e palavras-chave pode ter levado à exclusão de estudos cujo foco principal estivesse restrito ao corpo do texto, reduzindo a abrangência dos resultados. Esse procedimento, ainda que recorrente na literatura, representa uma limitação metodológica, pois estudos potencialmente relevantes podem não ter sido identificados. Recomenda-se, portanto, que pesquisas futuras reconheçam explicitamente essa limitação e incorporem estratégias complementares, como a triagem de texto completo em uma amostra ampliada ou o uso de técnicas de *Processamento de Linguagem Natural* (PLN) para capturar artigos que escapem às buscas tradicionais.

Outra ameaça diz respeito aos **procedimentos de inclusão e exclusão dos estudos**, que foram conduzidos de forma independente por dois avaliadores, contribuindo para reduzir vieses individuais. Entretanto, a **ausência de um terceiro revisor neutro** e a resolução das divergências somente por consenso configuram uma limitação metodológica. Essa prática, embora comum em mapeamentos sistemáticos, pode ter gerado decisões ambíguas e comprometido parcialmente a imparcialidade do processo de seleção. Para mitigar esse risco em trabalhos futuros, recomenda-se formalizar um protocolo explícito de resolução de divergências, contemplando, idealmente, a participação de um terceiro avaliador independente capaz de arbitrar eventuais discordâncias e fortalecer a robustez do processo de seleção dos estudos primários.

Além disso, foi identificado um risco relacionado à **extração e categorização dos dados**, realizadas manualmente e sem o respaldo de um processo sistemático de validação das categorias estabelecidas. Essa ausência de validação representa um potencial **viés interpretativo**, uma vez que as decisões dos revisores podem refletir julgamentos subjetivos. Para reduzir esse tipo de limitação, recomenda-se implementar um processo de validação sistemática, envolvendo, por exemplo, a extração independente de dados por dois revisores em uma subamostra de estudos, seguida da análise de concordância avaliadores, utilizando métricas como o coeficiente Kappa de Cohen. Além disso, o emprego de métodos formais para a criação e validação das categorias emergentes fortaleceria a consistência e a confiabilidade das análises conduzidas.

Outra limitação importante diz respeito à **abrangência da estratégia de busca**. Apesar de a busca ter sido conduzida de maneira fundamentada e revisada por especialistas, sua aplicação restrita a títulos, resumos e palavras-chave pode ter levado à exclusão de estudos relevantes cujos principais achados estavam contidos somente no corpo do texto. Essa limitação é reconhecida e deve ser considerada uma potencial ameaça à validade dos resultados. A inclusão

de abordagens complementares, como a revisão manual de uma amostra ampliada ou o uso de técnicas automáticas de varredura de texto, pode mitigar esse problema em estudos futuros.

Por fim, destaca-se a ameaça relacionada à **generalização dos resultados**. Por se tratar de um mapeamento sistemático, o estudo fornece uma visão ampla do estado da arte, mas não oferece evidências empíricas diretas sobre a eficácia das estratégias identificadas. A **inexistência de avaliações empíricas diretas** limita a capacidade de extrapolação dos resultados e impede conclusões robustas sobre a aplicabilidade prática das estratégias analisadas. Essa limitação é inerente ao delineamento metodológico adotado, que visa mapear o campo de pesquisa e não avaliar quantitativamente o impacto das soluções. Ainda assim, é importante reforçar essa limitação nas seções de discussão e considerações finais, destacando que os resultados apresentados funcionam como um alicerce para revisões sistemáticas futuras ou investigações empíricas voltadas à análise de eficácia e aplicação prática das estratégias. Nesse contexto, a proposta de expandir a classificação em direção à construção de um repositório de *benchmarks* representa um avanço metodológico relevante e um passo promissor na direção de estudos mais avaliativos e comparativos.

Em síntese, as ameaças à validade aqui identificadas relacionadas à **seleção dos estudos**, **procedimentos de inclusão e exclusão**, **extração e categorização de dados**, **clareza conceitual**, **estratégia de busca**, **qualidade textual** e **generalização dos resultados** não invalidam as conclusões do mapeamento, mas devem ser reconhecidas e tratadas de forma explícita. O reconhecimento dessas limitações contribui para a transparência científica e orienta aprimoramentos metodológicos em trabalhos subsequentes, reforçando a confiabilidade e a reprodutibilidade de futuras pesquisas na área.

6 LACUNAS IDENTIFICADAS E PERSPECTIVAS FUTURAS

Apesar dos avanços na computação em nuvem, a literatura ainda apresenta lacunas que limitam a generalização e a maturidade das soluções, principalmente na validação empírica, modelagem de cenários complexos, adaptação a ambientes heterogêneos e orquestração *multicloud* [6, 61]. Entre as estratégias promissoras, destacam-se abordagens baseadas em Inteligência Artificial e *Machine Learning*, fortemente associadas a termos como *neural networks*, *prediction*, *scheduling* e *resource allocation* [103]. Apesar de recorrentes, essas técnicas carecem de validação em ambientes reais, integração com práticas de FinOps e aplicação consistente em cenários *multicloud*, representando oportunidades relevantes para pesquisa [97]. A previsão e simulação de custos, vinculada a termos como *price prediction* e *cost management*, também se mostra promissora, mas apresenta baixa maturidade devido à escassez de modelos validados com dados reais e integração limitada a mecanismos automáticos de decisão [77].

A otimização de armazenamento por ciclo de vida e *tiering* engloba técnicas como migração entre classes, compressão, deduplicação e políticas automáticas de retenção e *caching* [120]. Contudo, faltam estudos sobre *trade-offs* custo–QoS, cobertura de cenários *multicloud* e exploração do *edge storage*. Na orquestração de contêineres em ambientes heterogêneos, predominam soluções para escalonamento de clusters, consolidação de cargas e *checkpoint/migração*, mas persistem desafios como a ausência de modelagem precisa do atraso de provisionamento e subutilização de aceleradores [126]. No paradigma *serverless*, as estratégias buscam reduzir custos via *function fusion/placement* e otimização de parâmetros, mas enfrentam dificuldades para equilibrar custo e latência, preservar modularidade e expandir para cenários *multicloud* [59].

Quatro pontos transversais sintetizam os desafios e direções futuras: (i) previsão e automação como pilares da economia, mas com escassez de *datasets* extensos e validação real [92]; (ii) governança do ciclo de vida de dados em estágio inicial [120]; (iii) integração entre modelos de precificação e previsão de demanda [29]; e (iv) carência de validação em produção, sobretudo em ambientes com hardware heterogêneo, *edge computing* e *serverless* em larga escala [59]. Essas lacunas reforçam a necessidade de estudos que unam modelagem precisa, integração de tecnologias emergentes e validação robusta em cenários reais, permitindo avanços significativos na eficiência e sustentabilidade econômica da computação em nuvem.

7 CONSIDERAÇÕES FINAIS

Este mapeamento sistemático apresentou uma análise detalhada das estratégias de redução de custos em computação em nuvem a partir da perspectiva do usuário, agrupando e classificando 82 estudos primários publicados no período de 2018 a 2024. As abordagens identificadas abrangem desde técnicas estabelecidas, como instâncias spot e reservadas, escalonamento automático, alocação otimizada de recursos e agendamento de tarefas, até práticas especializadas, incluindo otimização de contêineres, arquiteturas serverless, compressão e deduplicação de dados, políticas de ciclo de vida de armazenamento, uso de CDNs e balanceamento inteligente de carga.

A análise revelou predominância de estudos voltados à camada IaaS, especialmente para a otimização de máquinas virtuais, refletindo o controle granular e o impacto financeiro direto que essa camada oferece. Estratégias para armazenamento e rede apareceram em menor proporção, embora desempenhem papel relevante em cenários específicos. Identificou-se também forte concentração de trabalhos sobre nuvem pública, com menor atenção à *multicloud*, híbrida e praticamente ausência de estudos dedicados exclusivamente à nuvem privada.

A principal contribuição deste trabalho é fornecer uma visão abrangente e consolidada das práticas de redução de custos em nuvem, servindo de referência para pesquisadores, engenheiros e gestores. Para estudos futuros, recomenda-se expandir a classificação com detalhes sobre algoritmos, métricas, parâmetros de configuração e condições experimentais, visando aumentar a precisão na identificação de abordagens eficazes em diferentes contextos. Essa ampliação poderá fomentar a criação de um repositório padronizado de *benchmarks*, que viabilize comparações objetivas e incentive o aprimoramento contínuo das práticas de redução de custos na computação em nuvem.

REFERÊNCIAS

[1] Ehab Nabiel Al-Khanak, Sai Peck Lee, Saif Ur Rehman Khan, Navid Behboodian, Osamah Ibrahim Khalaf, Alexander Verbraeck, and Hans van Lint. 2021. A heuristics-based cost model for scientific workflow scheduling in cloud. 67, 3 (2021), 3265–3282. doi:10.32604/cmc.2021.015409

[2] Mana Saleh Al Reshan, Darakhshan Syed, Noman Islam, Asadullah Shaikh, Mohammed Hamdi, Mohamed A. Elmagzoub, Ghulam Muhammad, and Kashif Hussain Talpur. 2023. A Fast Converging and Globally Optimized Approach for Load Balancing in Cloud Computing. *IEEE Access* 11, February (2023), 11390–11404. doi:10.1109/ACCESS.2023.3241279

[3] Batool Alkaddah and Anjali Agarwal. 2022. Evaluating Amazon EC2 Spot Price Prediction Models Using Regression Error Characteristic Curve. *2022 7th International Conference on Fog and Mobile Edge Computing, FMEC 2022* (2022), 1–8. doi:10.1109/FMEC57183.2022.10062720

[4] Ehab Nabiel Alkhanak and Sai Peck Lee. 2018. A hyper-heuristic cost optimisation approach for Scientific Workflow Scheduling in cloud computing. *Future Gener. Comput. Syst.* 86 (2018), 480–506. doi:10.1016/j.future.2018.03.055

[5] Mohammed F. Alomari, Moamin A. Mahmoud, Niayesh Gharaei, Samer Mohammed Rasool, and Riyam A. Hasan. 2024. Optimizing Cloud Storage Costs: Introducing the Pre-Evaluation-Based Cost Optimization (PECSCO) Mechanism. *ICSINTESA 2024 - 2024 4th International Conference of Science and Information Technology in Smart Administration: The Collaboration of Smart Technology and Good Governance for Sustainable Development Goals* 1, 1 (2024), 564–569. doi:10.1109/ICSINTESA62455.2024.10748165

[6] Fahad Alshammari and Xiaohui Li. 2025. AI-driven cost optimization in cloud computing: A systematic review and future research directions. *J. Cloud Comput.* 14, 1 (2025), 1–22.

[7] Pradeep Ambati, Noman Bashir, David Irwin, Mohammad Hajiesmaili, and Prashant Shenoy. 2020. Hedge Your Bets: Optimizing Long-term Cloud Costs by Mixing VM Purchasing Options. *Proceedings - 2020 IEEE International Conference on Cloud Engineering, IC2E 2020* (2020), 105–115. doi:10.1109/IC2E48712.2020.00018

[8] Mohammed Amoon, Nirmeen El-Bahnasawy, and Mai ElKazaz. 2019. An efficient cost-based algorithm for scheduling workflow tasks in cloud computing systems. 31, 5 (2019), 1353–1363. doi:10.1007/s00521-018-3610-2

[9] David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, and James J. Cochran. 2020. *Statistics for Business and Economics* (13 ed.). Cengage Learning, Boston, MA.

[10] Jose Pergentino Araujo Neto, Donald M. Pianto, and Celia G. Ralha. 2018. An agent-based fog computing architecture for resilience on amazon EC2 spot instances. *Proceedings - 2018 Brazilian Conference on Intelligent Systems, BRACIS 2018* Cic (2018), 360–365. doi:10.1109/BRACIS.2018.00069

[11] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy H. Katz, Andrew Konwinski, Gunho Lee, David A. Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia. 2009. *Above the Clouds: A Berkeley View of Cloud Computing.* Technical Report UCB/EECS-2009-28. EECS Department, University of California, Berkeley, Berkeley, CA, USA. 1–25 pages. https://www2.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html

[12] Sugunakumar Arunan, Gayashan Amarasinghe, and Indika Perera. 2023. Cost-optimized scheduling for Microservices in Kubernetes. *Proceedings of the International Conference on Cloud Computing Technology and Science, CloudCom* (2023), 131–138. doi:10.1109/CloudCom59040.2023.00032

[13] S. Sajitha Banu and S. R. Balasundaram. 2021. Cost Optimization for Dynamic Content Delivery in Cloud-Based Content Delivery Network. 14, 4 (2021), 18–32. doi:10.4018/jitr.2021100102

[14] Sarah B. Basahel and Mohammad Yamin. 2022. A Novel Genetic Algorithm for Efficient Task Scheduling in Cloud Environment. *Proceedings of the 2022 9th International Conference on Computing for Sustainable Global Development, INDIACom 2022* (2022), 30–34. doi:10.23919/INDIACom54597.2022.9763230

[15] Manoj Bhoyar. 2025. AI-driven cloud optimization: Leveraging machine learning for dynamic resource allocation. *World J. Adv. Eng. Technol. Sci.* 15, 2 (2025), 877–884.

[16] Rajkumar Buyya, Satish Narayana Srirama, Giuliano Casale, Rodrigo Calheiros, Yogesh Simmhan, Blesson Varghese, Erol Gelenbe, Bahman Javadi, Luis Miguel Vaquero, Marco A. S. Netto, Adel Nadjaran Toosi, Maria Alejandra Rodriguez, Ignacio M. Llorente, Sabrina De Capitani di Vimercati, Pierangela Samarati, Dejan Milojicic, Carlos Varela, Rami Bahsoon, Marcos Dias de Assuncao, Omer Rana, Wanlei Zhou, Hai Jin, Wolfgang Gentzsch, Albert Y. Zomaya, and Haiying Shen. 2019. A Manifesto for Future Generation Cloud Computing: Research Directions for the Next Decade. *ACM Comput. Surv.* 51, 5 (2019), 105:1–105:38. doi:10.1145/3241737

[17] Rajkumar Buyya, Christian Vecchiola, and S. Thamarai Selvi. 2013. *Mastering Cloud Computing: Foundations and Applications Programming.* McGraw Hill Education, New York, NY, USA.

[18] Simon Caton, Matt Baughman, Christian Haas, Ryan Chard, Ian Foster, and Kyle Chard. 2022. Assessing the Current State of AWS Spot Market Forecastability. *Proceedings of SuperCompCloud 2022: 6th International Workshop on Interoperability of Supercomputing and Cloud Technologies, Held in conjunction with SC 2022: The International Conference for High Performance Computing, Networking, Storage and Analysis* 1, 1 (2022), 8–15. doi:10.1109/SuperCompCloud56703.2022.00007

[19] D. Chaudhary and Bijendra Kumar. 2019. Cost optimized Hybrid Genetic-Gravitational Search Algorithm for load scheduling in Cloud Computing. 83 (2019), 105627. doi:10.1016/j.asoc.2019.105627

[20] Junjie Chen and Hongjun Li. 2020. A two-phase cloud resource provisioning algorithm for cost optimization. 2020 (2020). doi:10.1155/2020/1310237

[21] Weihong Chen and Weichu Xiao. 2019. Cost-Efficient task scheduling for parallel applications on heterogeneous cloud environment. *Proceedings - 21st IEEE International Conference on High Performance Computing and Communications, 17th IEEE International Conference on Smart City and 5th IEEE International Conference on Data Science and Systems, HPCC/SmartCity/DSS 2019* (2019), 1651–1657. doi:10.1109/HPCC/SmartCity/DSS.2019.00226

[22] Mohan Baruwal Chhetri, Abdur Rahim Mohammad Forkan, Quoc Bao Vo, Surya Nepal, and Ryszard Kowalczyk. 2019. Towards risk-aware cost-optimal resource allocation for cloud applications. *Proceedings - 2019 IEEE International Conference on Services Computing, SCC 2019 - Part of the 2019 IEEE World Congress on Services* (2019), 210–214. doi:10.1109/SCC.2019.00043

[23] Yu Ting Chou, Shih Jui Liu, Tzu Chuan Wu, Chia Lin Wu, Chun We Tsai, and Ming Chao Chiang. 2018. An Effective Algorithm for Cloud Workflow Scheduling. *Proceedings - 2018 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2018* (2018), 3603–3608. doi:10.1109/SMC.2018.00609

[24] Louis Cohen, Lawrence Manion, and Keith Morrison. 2018. *Research Methods in Education* (8 ed.). Routledge, New York, NY. doi:10.4324/9781315456539

[25] John W. Creswell and J. David Creswell. 2017. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches.* Sage Publications, Thousand Oaks, CA.

[26] Leandro Costa Da Silva, Robson De Medeiros, and Nelson Rosa. 2023. COSTA: A cost-driven solution for migrating applications in multi-cloud environments. *Proceedings of the ACM Symposium on Applied Computing* (2023), 57–63. doi:10.1145/3555776.3577718

[27] Mustafa Daraghmeh, Anjali Agarwal, and Yaser Jararweh. 2023. Regression-Based Approach for Proactive Predictive Modeling of Efficient Cloud Cost Estimation. *2023 10th International Conference on Software Defined Systems, SDS 2023* (2023), 65–72. doi:10.1109/SDS59856.2023.10329194

[28] Himansu Das, Ajay Kumar Jena, J. Chandrakant Badajena, Chittaranjan Pradhan, and R. K. Barik. 2018. *Resource allocation in cooperative cloud environments.* Vol. 710. Springer Singapore. 825–841 pages. doi:10.1007/978-981-10-7871-2_79

[29] N. Deochake. 2024. Cloud cost optimization: A comprehensive review of strategies and case studies. *arXiv preprint arXiv:2307.12479* (2024).

[30] Saurabh Deochake. 2023. Cloud Cost Optimization: A Comprehensive Review of Strategies and Case Studies. Placeholder Publisher. https://ssrn.com/abstract=4519171.

[31] Jay L. Devore, Nicholas R. Farnum, and Jimmy A. Doi. 2018. *Applied Statistics for Engineers and Scientists* (3 ed.). Cengage Learning, Boston, MA.

[32] B. Dhayanandan and R. Rajeev. 2024. Cloud service price prediction using Machine Learning Algorithm with API in the case of Amazon Web Services and Microsoft Azure. *2024 International Conference on Intelligent Systems for Cybersecurity, ISCS 2024* (2024), 1–6. doi:10.1109/ISCS61804.2024.10581146

[33] Jose Luis Diaz, Joaquin Entrialgo, Javier Garcia, Manuel Garcia, and Daniel F. Garcia. 2021. Analysis of the Influence of Per-Second Billing on Virtual Machine Allocation Costs in Public Clouds. *IEEE Transactions on Services Computing* 14, 6 (2021), 1690–1701. doi:10.1109/TSC.2019.2909896

[34] Jose Luis Diaz, Javier Garcia, Joaquin Entrialgo, Manuel Garcia, and Daniel F. Garcia. 2020. Joint optimization of the cost of computation and virtual machine image storage in cloud infrastructure. *2020 International Symposium on Performance Evaluation of Computer and Telecommunication Systems, SPECTS 2020 - Proceedings* (2020).

[35] S. M.Reza Dibaj, Ali Miri, and Seyed Akbar Mostafavi. 2020. A cloud dynamic online double auction mechanism (DODAM) for sustainable pricing. *Telecommun. Syst.* 75, 4 (2020), 461–480. doi:10.1007/s11235-020-00688-4

[36] Quan Ding, Bo Tang, Prakash Manden, and Jin Ren. 2018. A learning-based cost management system for cloud computing. *2018 IEEE 8th Annual Computing and Communication Workshop and Conference, CCWC 2018* 2018-January (2018), 362–367. doi:10.1109/CCWC.2018.8301738

[37] Tore Dyba and Torgeir Dingsoyr. 2008. Empirical Studies of Agile Software Development: A Systematic Review. 50, 9-10 (2008), 833–859.

[38] Tarek Elgamal, Atul Sandur, Klara Nahrstedt, and Gul Agha. 2018. Optimizing cost of serverless computing through function fusion and placement. *Proceedings - 2018 3rd ACM/IEEE Symposium on Edge Computing, SEC 2018* (2018), 300–312. doi:10.1109/SEC.2018.00029

[39] Bugingo Emmanuel, Yingsheng Qin, Juntao Wang, Defu Zhang, and Wei Zheng. 2018. Cost optimization heuristics for deadline constrained workflow scheduling on clouds and their comparative evaluation. *Concurrency Comput.: Pract. Exper.* 30, 20 (2018), 1–14. doi:10.1002/cpe.4762

[40] K. Anders Ericsson and Herbert A. Simon. 1993. *Protocol Analysis: Verbal Reports as Data.* MIT Press, Cambridge, MA.

[41] Abdelkarim Erradi and Yaser Mansouri. 2020. Online cost optimization algorithms for tiered cloud storage services. 160 (2020), 110457. doi:10.1016/j.jss.2019.110457

[42] FinOps Foundation. 2025. *Calculating Container Costs.* Technical Report. FinOps Foundation. https://www.finops.org/wg/calculating-container-costs/ Focuses

on cost allocation, visibility, and optimization techniques in containerized environments such as Kubernetes. Relevant to compute cost optimization..

[43] FinOps Foundation. 2025. *How to Optimize Cloud Usage*. Technical Report. FinOps Foundation. https://www.finops.org/wg/how-to-optimize-cloud-usage/ Includes practices for compute (right-sizing, autoscaling, spot instances, serverless), storage (snapshot cleanup, lifecycle policies, storage tiering), and network (data transfer optimization, CDN, caching)..

[44] FinOps Foundation. 2025. *Usage Optimization Opportunities Library*. Technical Report. FinOps Foundation. https://www.finops.org/wg/workload-optimization/ Presents practical use cases such as identifying unused snapshots, aborting incomplete multipart uploads, and enabling serverless tiers for databases..

[45] Flexera. 2025. Flexera 2025 State of the Cloud Report. https://info.flexera.com/CM-REPORT-State-of-the-Cloud#CM-REPORT-State-of-the-Cloud-2025. 14ª edição anual. Publicado em 19 de março de 2025; acesso em: 07 ago. 2025.

[46] George Fragiadakis, Evangelia Filiopoulou, Christos Michalakelis, Thomas Kamalakis, and Mara Nikolaidou. 2023. Applying Machine Learning in Cloud Service Price Prediction: The Case of Amazon IaaS. 15, 8 (2023). doi:10.3390/fi15080277

[47] Prerna Gaba, Himanshu, Preetitanya, and Yatin Gupta. 2023. Unlocking Efficiency - Multidimensional Cost Optimization Strategies for Cloud Infrastructure in Small and Medium-Sized Organizations. *2nd International Conference on Automation, Computing and Renewable Systems, ICACRS 2023 - Proceedings* (2023), 463–470. doi:10.1109/ICACRS58579.2023.10404455

[48] Javier Garcia, Joaquin Entrialgo, Jose Luis Diaz, Manuel Garcia, and Daniel F. Garcia. 2019. Influence of the trace resolution and length in the cost optimization process in cloud computing. *Proceedings of the 2019 International Symposium on Performance Evaluation of Computer and Telecommunication Systems, SPECTS 2019 - Part of SummerSim 2019 Multiconference* XX, YY (2019). doi:10.23919/SPECTS.2019.8823508

[49] Gareth George, Rich Wolski, Chandra Krintz, and John Brevik. 2019. Analyzing AWS spot instance pricing. *Proceedings - 2019 IEEE International Conference on Cloud Engineering, IC2E 2019* (2019), 222–228. doi:10.1109/IC2E.2019.00036

[50] Muhammad Hamza, Muhammad Azeem Akbar, and Rafael Capilla. 2023. Understanding cost dynamics of serverless computing: An empirical study. In *International Conference on Software Business*. Springer-Verlag, Lappeenranta, Finland, 456–470.

[51] Hassan B Hassan, Saman A Barakat, and Qusay I Sarhan. 2021. Survey on serverless computing. *J. Cloud Comput.* 10, 1 (2021), 39.

[52] Marie C. Hoepfl. 1997. Choosing Qualitative Research: A Primer for Technology Education Researchers. 9, 1 (1997), 47–63.

[53] Tarun Jain and Jishnu Hazra. 2019. "On-demand" pricing and capacity management in cloud computing. 18, 3 (2019), 228–246. doi:10.1057/s41272-018-0146-0

[54] S Jayalakshmi et al. 2021. Predictive scaling for elastic compute resources on public cloud utilizing deep learning based long short-term memory. *Int. J. Adv. Comput. Sci. Appl.* 12, 10 (2021), 1–20.

[55] Steffen Kächele, Christian Spann, Franz J. Hauck, and Jörg Domaschka. 2013. Beyond IaaS and PaaS: An Extended Cloud Taxonomy for Computation, Storage and Networking. In *Proceedings of the 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing (UCC '13)*. IEEE Computer Society, Washington, DC, USA, 75–82. doi:10.1109/UCC.2013.28

[56] Ian A. Kash, Peter Key, and Warut Suksompong. 2019. Simple pricing schemes for the cloud. *ACM Transactions on Economics and Computation* 7, 2 (2019), 1–27. doi:10.1145/3327973

[57] Deepak Kaul. 2019. Optimizing Resource Allocation in Multi-Cloud Environments with Artificial Intelligence: Balancing Cost, Performance, and Security. *Journal of Information and Computer Technology Education)* January 2019 (2019). https://questsquare.org/index.php

[58] Akif Quddus Khan, Mihhail Matskin, Radu Prodan, Christoph Bussler, Dumitru Roman, and Ahmet Soylu. 2024. *Cloud storage cost: a taxonomy and survey*. Vol. 27. Placeholder Publisher. 1–54 pages. doi:10.1007/s11280-024-01273-4

[59] Akif Quddus Khan, Mihhail Matskin, Radu Prodan, Christoph Bussler, Dumitru Roman, and Ahmet Soylu. 2024. Cost modelling and optimisation for cloud: a graph-based approach. *J. Cloud Comput.* 13, 1 (2024), 147.

[60] Khalid S Khan, Gerben Ter Riet, Julie Glanville, Amanda J Sowden, Jos Kleijnen, et al. 2001. *Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews*. Number 4 (2n in CRD Report. NHS Centre for Reviews and Dissemination, York.

[61] Jae Kim and Minsoo Park. 2025. Holistic multi-objective container orchestration for heterogeneous cloud clusters. *IEEE Trans. Cloud Comput.* (2025). Early Access.

[62] Barbara Kitchenham. 2007. Guidelines for performing Systematic Literature Reviews in Software Engineering. *EBSE Technical Report* 2, 1 (2007), 1–57.

[63] Barbara Kitchenham, O. Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. 2009. Systematic Literature Reviews in Software Engineering – A Systematic Literature Review. 51, 1 (2009), 7–15.

[64] Barbara A. Kitchenham. 2004. Systematic Reviews. In *Proceedings of the 10th International Symposium on Software Metrics*. IEEE, IEEE, Chicago, IL, xii–xii.

[65] Barbara Ann Kitchenham, David Budgen, and Pearl Brereton. 2015. *Evidence-Based Software Engineering and Systematic Reviews*. Vol. 4. CRC Press, Boca Raton, FL. doi:10.1201/b19467

[66] Dawei Kong, Shijun Liu, and Li Pan. 2021. Amazon Spot Instance Price Prediction with GRU Network. *Proceedings of the 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2021* (2021), 31–36. doi:10.1109/CSCWD49262.2021.9437881

[67] Dinesh Kumar, Gaurav Baranwal, Zahid Raza, and Deo Prakash Vidyarthi. 2018. A Survey on Spot Pricing in Cloud Computing. 26, 4 (2018), 809–856. doi:10.1007/s10922-017-9444-x

[68] K. Dinesh Kumar and E. Umamaheswari. 2018. Prediction methods for effective resource provisioning in cloud computing: A survey. *Multiagent Grid Syst.* 14, 3 (2018), 283–305. doi:10.3233/MGS-180292

[69] Eva Maria Lakatos and Marina de Andrade Marconi. 2003. *Fundamentos de Metodologia Científica* (5 ed.). Atlas, São Paulo.

[70] Chunlin Li, Jingpan Bai, and Youlong Luo. 2020. *Efficient resource scaling based on load fluctuation in edge-cloud computing environment*. Vol. 76. Springer US. 6994–7025 pages. doi:10.1007/s11227-019-03134-8

[71] Fang Li, Gang Wu, Jianhua Lu, Mingye Jin, Haitao An, and Junxiong Lin. 2022. SmartCMP: A Cloud Cost Optimization Governance Practice of Smart Cloud Management Platform. *Proceedings - 2022 IEEE 7th International Conference on Smart Cloud, SmartCloud 2022* X, Y (2022), 171–176. doi:10.1109/SmartCloud55982.2022.00034

[72] Zhe Li, Yusong Tan, Bao Li, Jianfeng Zhang, and Xiaochuan Wang. 2021. A Survey of Cost Optimization in Serverless Cloud Computing. *IOP Conference Series: Earth and Environmental Science* 1802, 3 (2021). doi:10.1088/1742-6596/1802/3/032070

[73] Zengpeng Li, Huiqun Yu, and Guisheng Fan. 2023. Cost-effective approaches for deadline-constrained workflow scheduling in clouds. *J. Supercomput.* 79, 7 (2023), 7484–7512. doi:10.1007/s11227-022-04962-x

[74] Changyuan Lin and Hamzeh Khazaei. 2021. Modeling and Optimization of Performance and Cost of Serverless Applications. *IEEE Transactions on Parallel and Distributed Systems* 32, 3 (2021), 615–632. doi:10.1109/TPDS.2020.3028841

[75] Mingyu Liu, Li Pan, and Shijun Liu. 2021. Keep Hot or Go Cold: A Randomized Online Migration Algorithm for Cost Optimization in STaaS Clouds. *IEEE Trans. Netw. Serv. Manag.* 18, 4 (2021), 4563–4575. doi:10.1109/TNSM.2021.3096533

[76] Mingyu Liu, Li Pan, and Shijun Liu. 2023. Cost Optimization for Cloud Storage from User Perspectives: Recent Advances, Taxonomy, and Survey. *ACM Comput. Surv.* 55, 13s (2023), 1–20. doi:10.1145/3582883

[77] Maria Lopez and Pedro Fernandes. 2024. Edge-aware storage tiering for cost and latency optimization in hybrid clouds. *Future Gener. Comput. Syst.* 158 (2024), 299–315.

[78] Sharmistha Mandal, Giridhar Maji, Sunirmal Khatua, and Rajib K. Das. 2023. Cost Minimizing Reservation and Scheduling Algorithms for Public Clouds. *IEEE Trans. Cloud Comput.* 11, 2 (2023), 1365–1380. doi:10.1109/TCC.2021.3133464

[79] Yaser Mansouri and Abdelkarim Erradi. 2018. Cost Optimization Algorithms for Hot and Cool Tiers Cloud Storage Services. *IEEE International Conference on Cloud Computing, CLOUD* 2018-July (2018), 622–629. doi:10.1109/CLOUD.2018.00086

[80] Marina de Andrade Marconi and Eva Maria Lakatos. 2012. Técnicas de pesquisa: Planejamento e execução de pesquisas, amostragens e técnicas de pesquisa, elaboração, análise e interpretação de dados. In *Técnicas de pesquisa: planejamento e execução de pesquisa; amostragens e técnicas de pesquisa; elaboração, análise e interpretação de dados* (7 ed.). Atlas, São Paulo, 277–277.

[81] Dan C. Marinescu. 2017. *Cloud Computing: Theory and Practice* (2nd ed.). Morgan Kaufmann, Boston, MA, USA.

[82] Peter Mell and Timothy Grance. 2011. *The NIST Definition of Cloud Computing*. Special Publication 800-145. National Institute of Standards and Technology, Gaithersburg, MD, USA. https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

[83] Peter Mell and Timothy Grance. 2011. *The NIST Definition of Cloud Computing*. NIST Special Publication 800-145. National Institute of Standards and Technology, Gaithersburg, MD, USA. 1–7 pages. doi:10.6028/NIST.SP.800-145

[84] Fanchao Meng, Qingran Ji, Dinahui Chu, and Xuequan Zhou. 2021. Modeling and Solution Algorithm of Virtual Machines Optimization Provision Problem for Application Deployment in Public Cloud. *19th IEEE International Symposium on Parallel and Distributed Processing with Applications, 11th IEEE International Conference on Big Data and Cloud Computing, 14th IEEE International Conference on Social Computing and Networking and 11th IEEE International Conference on Sustainable Computing and Communications, ISPA/BDCloud/SocialCom/SustainCom 2021* (2021), 1378–1385. doi:10.1109/ISPA-BDCloud-SocialCom-SustainCom52081.2021.00188

[85] Dimitar Mileski and Marjan Gusev. 2023. FinOps in Cloud-Native Near Real-Time Serverless Streaming Solutions. *2023 31st Telecommunications Forum, TELFOR 2023 - Proceedings* (2023), 1–4. doi:10.1109/TELFOR59449.2023.10372626

[86] Ashish Kumar Mishra, Brajesh Kumar Umrao, and Dharmendra K. Yadav. 2018. A survey on optimal utilization of preemptible VM instances in cloud computing. *J. Supercomput.* 74, 11 (2018), 5980–6032. doi:10.1007/s11227-018-2509-0

[87] David A. Monge, Elina Pacini, Cristian Mateos, and Carlos García Garino. 2018. Meta-heuristic based autoscaling of cloud-based parameter sweep experiments with unreliable virtual machines instances. *Comput. Electr. Eng.* 69 (2018), 364–377. doi:10.1016/j.compeleceng.2017.12.007

[88] Koyel Mukherjee, Raunak Shah, Shiv Saini, Karanpreet Singh, Khushi, Harsh Kesarwani, Kavya Barnwal, and Ayush Chauhan. 2023. Towards Optimizing Storage Costs on the Cloud. *Proceedings - International Conference on Data Engineering* 2023-April (2023), 2919–2932. doi:10.1109/ICDE55515.2023.00223

[89] Piotr Nawrocki and Mateusz Smendowski. 2023. Long-Term Prediction of Cloud Resource Usage in High-Performance Computing. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 14074 LNCS (2023), 532–546. doi:10.1007/978-3-031-36021-3_53

[90] Piotr Nawrocki and Mateusz Smendowski. 2024. FinOps-driven optimization of cloud resource usage for high-performance computing using machine learning. *J. Comput. Sci.* 79, April (2024), 102292. doi:10.1016/j.jocs.2024.102292

[91] Piotr Nawrocki and Mateusz Smendowski. 2024. Optimization of the Use of Cloud Computing Resources Using Exploratory Data Analysis and Machine Learning. 14, 4 (2024), 287–308. doi:10.2478/jaiscr-2024-0016

[92] Piotr Nawrocki and Mateusz Smendowski. 2025. A Survey of Cloud Resource Consumption Optimization Methods. 23, 5 (2025), 5. doi:10.1007/s10723-024-09792-0

[93] Jose Pergentino A. Neto, Donald M. Pianto, and Célia Ghedini Ralha. 2018. A Prediction Approach to Define Checkpoint Intervals in Spot Instances. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10967 LNCS, June (2018), 84–93. doi:10.1007/978-3-319-94295-7_6

[94] Seyed Soroush Nezamdoust, Mohammad Ali Pourmina, and Farbod Razzazi. 2023. Optimal prediction of cloud spot instance price utilizing deep learning. *J. Supercomput.* 79, 7 (2023), 7626–7647. doi:10.1007/s11227-022-04970-x

[95] Patryk Osypanka and Piotr Nawrocki. 2022. Resource Usage Cost Optimization in Cloud Computing Using Machine Learning. *IEEE Trans. Cloud Comput.* 10, 3 (2022), 2079–2089. doi:10.1109/TCC.2020.3015769

[96] Jay Oza, Rishi More, Amit Maity, Gitesh Kambli, Chirag Maniyath, and Abhijit Patil. 2024. PRISM: Predictive Resource Inference and Spot Instance Management. *2024 3rd International Conference for Advancement in Technology, ICONAT 2024* XX, YY (2024), 1–6. doi:10.1109/ICONAT61936.2024.10774810

[97] Rajesh Patel and Arjun Singh. 2024. FinOps 2.0: Intelligent financial operations for multi-cloud environments. *ACM Comput. Surv.* 56, 4 (2024), 1–38.

[98] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. 2008. Systematic Mapping Studies in Software Engineering. In *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering (EASE)* (Italy). BCS Learning & Development, BCS, Swindon, GBR, 1–10.

[99] Mark Petticrew and Helen Roberts. 2008. *Systematic Reviews in the Social Sciences: A Practical Guide.* John Wiley & Sons, Chichester, UK.

[100] Sivakumar Ponnusamy and Mandar Khoje. 2024. Optimizing Cloud Costs with Machine Learning: Predictive Resource Scaling Strategies. *2024 5th International Conference on Innovative Trends in Information Technology, ICITIIT 2024* X, Y (2024), 1–8. doi:10.1109/ICITIIT61487.2024.10580717

[101] S. Ramamoorthy, G. Ravikumar, B. Saravana Balaji, S. Balakrishnan, and K. Venkatachalam. 2021. MCAMO: multi constraint aware multi-objective resource scheduling optimization technique for cloud infrastructure services. *J. Ambient Intell. Humaniz. Comput.* 12, 6 (2021), 5909–5916. doi:10.1007/s12652-020-02138-0

[102] Aishwarya Ramesh, Vishal Pradhan, and Hemraj Lamkuche. 2021. Understanding and Analysing Resource Utilization, Costing Strategies and Pricing Models in Cloud Computing. *J. Phys.: Conf. Ser.* 1964, 4 (2021), 042049. doi:10.1088/1742-6596/1964/4/042049

[103] Shweta Rani and Manoj Gupta. 2024. Predictive resource allocation using deep reinforcement learning for cloud cost efficiency. *IEEE Trans. Netw. Serv. Manag.* (2024). Early Access.

[104] Pradeep Singh Rawat, Priti Dimri, and Gyanendra Pal Saroha. 2020. Virtual machine allocation to the task using an optimization method in cloud computing environment. *Int. J. Inf. Technol.* 12, 2 (2020), 485–493. doi:10.1007/s41870-018-0242-9

[105] Thiago Reis, Mario Teixeira, Joao Almeida, and Anselmo Paiva. 2020. A Recommender for Resource Allocation in Compute Clouds Using Genetic Algorithms and SVR. *IEEE Latin America Transactions* 18, 6 (2020), 1049–1056. doi:10.1109/TLA.2020.9099682

[106] Dhruv Seth, Harshavardhan Nerella, Madhavi Najana, and Ayisha Tabbassum. 2024. Navigating the multi-cloud maze: benefits, challenges, and future trends. *Int. J. Glob. Innov. Solut.* (2024).

[107] Mark Shifrin, Roy Mitrany, Erez Biton, and Omer Gurewitz. 2022. VM Scaling and Load Balancing via Cost Optimal MDP Solution. *IEEE Trans. Cloud Comput.* 10, 3 (2022), 2219–2237. doi:10.1109/TCC.2020.3000956

[108] Vivek Kumar Singh, Shivendu Shivendu, and Kaushik Dutta. 2022. Spot instance similarity and substitution effect in cloud spot market. *Decis. Support Syst.* 159, November 2021 (2022), 113815. doi:10.1016/j.dss.2022.113815

[109] Mateusz Smendowski and Piotr Nawrocki. 2024. Optimizing multi-time series forecasting for enhanced cloud resource utilization based on machine learning. *Knowl.-Based Syst.* 304, June (2024), 112489. doi:10.1016/j.knosys.2024.112489

[110] Georgios Spanos and Lefteris Angelis. 2016. The Impact of Information Security Events to the Stock Market: A Systematic Literature Review. *Computers & Security* 58 (2016), 216–229. doi:10.1016/j.cose.2015.12.006

[111] Devesh Kumar Srivastava, Sumit Kumar Gupta, Pradeep Kumar Tiwari, and Manjiit Kaur. 2024. Resource Management on Cloud Computing Using Machine Learning. *Proceedings - International Conference on Computational Intelligence and Networks* XX, YY (2024), 1–6. doi:10.1109/CINE63708.2024.10881266

[112] Kavita Srivastava and Manisha Agarwal. 2024. Maximizing Cloud Resource Utility: Region-Adaptive Optimization via Machine Learning-Informed Spot Price Predictions. *Lecture Notes in Networks and Systems* 997 LNNS (2024), 449–459. doi:10.1007/978-981-97-3242-5_30

[113] J.R. Storment and M. Fuller. 2023. *Cloud FinOps: Collaborative, Real-Time Cloud Financial Management* (2nd ed.). O'Reilly Media, Inc., Sebastopol, CA. doi:10.5555/3587612

[114] Fuquan Sun, Zhenghao Lu, Jikui Pan, and Zijian Wang. 2021. A Cost Optimization Strategy for Workflow Scheduling in Cloud. *Proceedings of the 33rd Chinese Control and Decision Conference, CCDC 2021* 1, 1 (2021), 270–274. doi:10.1109/CCDC52312.2021.9601544

[115] Cristiano Costa Argemon Vieira, Luiz Fernando Bittencourt, Thiago Augusto Lopes Genez, Maycon Leone M. Peixoto, and Edmundo Roberto Mauro Madeira. 2024. RAaaS: Resource Allocation as a Service in multiple cloud providers. 221, February 2023 (2024), 103790. doi:10.1016/j.jnca.2023.103790

[116] Mira Vrbaski, Miodrag Bolic, and Shikharesh Majumdar. 2022. Multi-objective optimization for cloud provisioning: A case study in large-scale microservice notification applications. *Proceedings - 2022 International Conference on Future Internet of Things and Cloud, FiCloud 2022* XX, YY (2022), 190–198. doi:10.1109/FiCloud57274.2022.00033

[117] Danjing Wang, Huifang Li, Youwei Zhang, and Baihai Zhang. 2023. Gradient-Based Scheduler for Scientific Workflows in Cloud Computing. 27, 1 (2023), 64–73. doi:10.20965/jaciii.2023.p0064

[118] Caesar Wu, Rajkumar Buyya, and Kotagiri Ramamohanarao. 2019. Cloud Pricing Models: Taxonomy, Survey, and Interdisciplinary Challenges. *ACM Comput. Surv.* 52, 6 (2019), 108:1–108:36. doi:10.1145/3342103

[119] Mingyu Wu, Zeyu Mi, and Yubin Xia. 2020. A Survey on Serverless Computing and Its Implications for JointCloud Computing. In *2020 IEEE International Conference on Joint Cloud Computing*. IEEE, New York, NY, 94–101. doi:10.1109/JCC49151.2020.00023

[120] Y. Wu, R. Buyya, et al. 2024. Cost optimization for cloud storage from user perspectives: Recent advances, taxonomy, and survey. *World Wide Web* 27, 1 (2024), 45–78.

[121] Yongjie Xie, Li Pan, Shengsong Yang, and Shijun Liu. 2022. A Random Online Algorithm for Reselling Reserved IaaS Instances in Amazon's Cloud Marketplace. *IEEE Transactions on Network Science and Engineering* 9, 3 (2022), 1235–1244. doi:10.1109/TNSE.2021.3138932

[122] Shengsong Yang, Li Pan, and Shijun Liu. 2019. An online algorithm for selling your reserved IaaS instances in amazon EC2 marketplace. *Proceedings - 2019 IEEE International Conference on Web Services, ICWS 2019 - Part of the 2019 IEEE World Congress on Services* X, Y (2019), 296–303. doi:10.1109/ICWS.2019.00057

[123] Shengsong Yang, Li Pan, Qingyang Wang, and Shijun Liu. 2018. To sell or not to sell: Trading your reserved instances in amazon EC2 marketplace. *Proceedings - International Conference on Distributed Computing Systems* 2018-July, 1 (2018), 939–948. doi:10.1109/ICDCS.2018.00095

[124] Shengsong Yang, Li Pan, Qingyang Wang, Shijun Liu, and Shuo Zhang. 2018. Subscription or Pay-as-You-Go: Optimally Purchasing IaaS Instances in Public Clouds. In *Proceedings - 2018 IEEE International Conference on Web Services, ICWS 2018 - Part of the 2018 IEEE World Congress on Services*. IEEE, New York, NY, USA, 219–226. doi:10.1109/ICWS.2018.00035

[125] Rehmana Younis, Muhammad Aaqib Javed, Mansoor Iqbal, Khalid Munir, Muhammad Haris, and Saad Alahmari. 2024. A Comprehensive Analysis of Cloud Service Models: IaaS, PaaS and SaaS in the Context of Emerging Technologies and Trend. In *Proceedings of the 2024 International Conference on Electrical, Communication and Computer Engineering (ICECCE)*. IEEE, Kuala Lumpur, Malaysia, 1–6. doi:10.1109/ICECCE63537.2024.10823401

[126] H. Zhang et al. 2024. Intelligent orchestration for heterogeneous cloud environments. *Future Gener. Comput. Syst.* 157 (2024), 512–526.

[127] Peipei Zhou, Jiayi Sheng, Cody Hao Yu, Peng Wei, Jie Wang, Di Wu, and Jason Cong. 2021. MOCHA: Multinode cost optimization in heterogeneous clouds with accelerators. *FPGA 2021 - 2021 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays* 1, 1 (2021), 273–279. doi:10.1145/3431920.3439304

[128] Xiumin Zhou, Gongxuan Zhang, Jin Sun, Junlong Zhou, Tongquan Wei, and Shiyan Hu. 2019. Minimizing cost and makespan for workflow scheduling in cloud using fuzzy dominance sort based HEFT. *Future Gener. Comput. Syst.* 93 (2019), 278–289. doi:10.1016/j.future.2018.10.046