O Comitê Central da Nuvem: Um Estudo de Caso de Planejamento de Capacidade em Grandes Empresas de TI

Pedro Serey[1], **Caio Galvão**[1], **Thiago Emmanuel Pereira**[1], **Francisco Vilar Brasileiro**[1], **Gabriel Gomes**[1,2]

[1]Universidade Federal de Campina Grande
Unidade Acadêmica de Sistemas e Computação
Av. Aprígio Veloso, s/n, Campina Grande, PB, 58.429-900, Brasil

{pedro.serey, caio.galvao}@ccc.ufcg.edu.br

{temmanuel,fubica}@computacao.ufcg.edu.br

[2]VTEX
Centro Empresarial, Praia de Botafogo, 300 - 3º Andar - Botafogo
Rio de Janeiro - RJ, 22250-040, Brasil

gabriel.cardoso@vtex.com

Abstract. *Large companies often decentralize into sub-organizations to accelerate decision-making. For instance, when using IaaS, these sub-organizations might independently purchase cloud resources. While flexible, such a decentralized approach can be more expensive than a company-wide, centralized strategy, due to a wider variation in usage patterns of sub-organizations. To investigate these trade-offs within the scenario of a major IaaS consumer, we analyzed multiple years of cloud purchase data from a partner company to quantify potential cost savings under different organizational models. Our findings show that centralization yielded significant savings, potentially reducing IaaS costs by nearly* 9%*, translating into millions of dollars in savings in this case study.*

Resumo. *Grandes empresas descentralizam suas operações em suborganizações para acelerar a tomada de decisões. Por exemplo, as suborganizações podem adquirir recursos de provedores de IaaS de forma independente. Embora flexível, a abordagem descentralizada pode ser mais cara do que a centralizada, devido à maior variabilidade da demanda das suborganizações. Analisamos as compras de vários anos de uma empresa parceira, grande consumidora de IaaS, quantificando as economias obtidas sob diferentes modelos organizacionais. Observamos que a centralização pode gerar economias significativas: no caso em questão, temos quase* 9% *de redução de custos de IaaS, gerando uma economia de milhões de dólares.*

1. Introdução

Computação na nuvem possibilitou a compra de recursos sob demanda. Isso trouxe uma vantagem significativa em comparação aos tempos anteriores à nuvem, eliminando a necessidade de estimar o pico da demanda antes da aquisição e correr o risco de superestimar, resultando em altos custos e recursos ociosos, ou, no pior dos casos, subestimar o pico e enfrentar falta de recursos em momentos críticos. Embora a compra sob demanda

tenha seus benefícios, os provedores oferecem opções mais econômicas ao reservar recursos por períodos mais longos, que são bastante atrativos para grandes consumidores. Contudo, essas diferentes opções trazem de volta o mesmo problema de estimativa de demanda, revelando que o planejamento de capacidade nunca deixou de ser relevante.

Tipicamente, os provedores de IaaS adotam diversas estratégias para minimizar os riscos de superprovisionamento, como a definição de cotas de capacidade fixa para os usuários [Costa et al. 2013] e a criação de mercados de compra. Esses mercados oferecem diferentes trade-offs entre preço, tempo de compromisso, disponibilidade e obtenabilidade dos recursos [Carvalho et al. 2014]. Consequentemente, conhecer a demanda da empresa com precisão pode levar a reduções significativas de custos. Além disso, devido a algumas características de mercados específicos, como o AWS Savings Plans – onde os descontos são aplicados às reservas dentro de famílias específicas de instâncias AWS (que agrupam tipos de instâncias) – a estimativa de demanda deve considerar detalhes como tipos, famílias e padrões de uso ao longo do tempo das instâncias.

Embora informações precisas sobre a demanda sejam essenciais para realizar compras adequadas, essas informações frequentemente residem nos usuários finais – desenvolvedores, gerentes de produto e membros das equipes em geral – que interagem diretamente com a alocação e uso desses recursos em seu trabalho diário. Cullen e Perrewé [Cullen and Perrewé 1981] observaram que, em organizações centralizadas, a administração superior frequentemente carece do conhecimento especializado necessário para supervisionar diretamente áreas que requerem um alto grau de expertise técnica.

Organizações descentralizadas trazem a flexibilidade e a autonomia necessárias para as sub-organizações, permitindo que tomem decisões, comuniquem-se e lidem melhor com desafios inesperados, como afirmado por Englehardt e Simmons [Englehardt and Simmons 2002]. Contudo, a aquisição de recursos em nuvem no nível das sub-organizações pode levar a uma demanda fragmentada no contexto da empresa. Sendo assim, modelos centralizados podem evitar a desordem organizacional – como a fragmentação da demanda por IaaS – e alcançar vantagens de escala por meio da cooperação entre as sub-organizações, por exemplo, agregando as demandas e justificando as compras nos mercados que exigem compromissos de longo prazo. Portanto, embora perca autonomia em um ambiente como a aquisição de IaaS, que exige conhecimento especializado, esse modelo promove a cooperação entre sub-organizações, potencialmente alcançando redução de custos.

A questão central nesse contexto é quantificar a possível economia de custo na centralização das decisões de compra de recursos em nuvem. Considerando esse problema, analisamos um conjunto de dados de longa duração da demanda por recursos de computação na nuvem da VTEX, uma grande empresa de *e-commerce*, para entender sua organização. Com base na análise da estratégia de compra da empresa e utilizando um modelo de otimização para a compra de recursos na nuvem [Galvão et al. 2024], utilizamos uma amostra de três anos do conjunto de dados principal (janeiro/2021 a dezembro/2023) para avaliar as potenciais economias de custo na centralização das decisões de aluguel junto ao provedor de IaaS da empresa. Embora a literatura em gestão mostre que a conceitualização de Centralização/Descentralização é mais complexa do que um espectro que vai de totalmente descentralizado a totalmente centralizado [Cullen and Perrewé 1981], este estudo utilizou o conceito mais simples de espectro

linear para os experimentos realizados.

O restante deste artigo está organizado da seguinte forma: na Seção 2, apresentamos uma análise descritiva da estrutura organizacional da empresa estudada. Na Seção 3, explicamos a metodologia do estudo de caso, incluindo os experimentos realizados. Na Seção 4, apresentamos os resultados do estudo de caso para os cenários abordados na Seção 3. Por fim, na Seção 5, discutimos possíveis trabalhos futuros.

2. Análise Organizacional

Esta seção analisa a organização da empresa, incluindo como suas suborganizações são estruturadas. Inicialmente, é importante entender que essas suborganizações compõem o setor de engenharia da empresa. Os funcionários da empresa são divididos em times que desenvolvem produtos, os quais são agrupados, juntamente com times de produtos similares, em verticais. Essas suborganizações utilizam recursos na nuvem para operar serviços internos e externos, realizar experimentos, desenvolver soluções, entre outras atividades.

Embora essa estrutura organizacional exista há muito tempo na empresa, só mais recentemente foi possível associar, através de um serviço de auditoria, para cada máquina virtual alugada na nuvem, a identificação de time e vertical responsável pelo recurso alocado. Após analisar o conjunto de dados, selecionamos para o nosso estudo de caso os últimos três anos do conjunto de dados: 2021, 2022 e 2023. Esses anos correspondem ao período com a maior demanda mapeada em times e verticais (cerca de 78% da demanda total).

Um dos primeiros passos na análise do conjunto de dados da empresa foi quantificar as suborganizações e entender sua escala. Nos últimos três anos de dados registrados, quase 600 times únicos foram identificados, a maioria dos times está alinhada com uma das 13 verticais da empresa. Como dissemos, essas verticais representam aproximadamente 78% da demanda mapeada. Os 22% restantes da demanda incluem 266 times que não estão mapeados para nenhuma das verticais. Essa porção não mapeada mostra que os recursos podem ser alocados fora das verticais, seja por verticais emergentes ou times que desenvolvem produtos não relacionados a nenhuma delas e que, eventualmente, podem se juntar a uma das verticais.

Considerando essas informações, há decisões importantes a serem tomadas antes de iniciar o estudo de caso. Uma das primeiras decisões foi não considerar o nível de "time" como uma suborganização. Tomamos essa decisão porque, como as demandas dos times podem ser muito pequenas, a aquisição de recursos nesse nível seria trivialmente subótima em relação à demanda unificada no modelo centralizado.

3. Metodologia

Nesta seção, apresentamos uma visão geral da metodologia utilizada para realizar o estudo de caso. O ponto-chave da metodologia é comparar o custo de aluguel de recursos na nuvem para os diferentes cenários organizacionais. Com base nas informações sobre a organização da empresa apresentadas na Seção 2, geramos as demandas das verticais e decidimos a escolha de aluguel na nuvem para cada uma individualmente, assim como para a demanda completa. Com os dados resultantes, calculamos os custos com base nas

informações de preços dos tipos de instância alugados e os comparamos, quantificando os descontos potenciais.

O conjunto de dados que consideramos nesse estudo abrange 10 anos de uso de instâncias na nuvem pela VTEX. Os dados consistem em registros de alocação de máquinas virtuais em intervalos de uma hora. Esses registros contêm as seguintes informações: 1) **Tipo de Instância**, descritor disponibilizado pelo provedor IaaS, que permite identificar a capacidade da instância, incluindo o número de CPUs virtuais e a quantidade de RAM; 2) **Timestamp**, o horário de início do intervalo de uso da máquina virtual; 3) **Mercado**, indica o mercado AWS do contrato associado à máquina virtual alocada; 4) **Tag da vertical**, identificador da vertical responsável pela máquina virtual alocada; e 5) **Tag do time**, identificador do time responsável pela máquina virtual alocada.

Como discutimos anteriormente, nem todos os registros de alocação estão mapeados para uma vertical. Duas abordagens podem ser utilizadas para resolver esse problema: considerar a *demanda não mapeada* como uma demanda adicional a ser comprada, junto à demanda das verticais mapeadas, ou remover os registros não mapeados da demanda. Após analisar a *demanda não mapeada*, percebemos que o comportamento dessa demanda não era muito diferente do das verticais. Com isso, conduzimos o estudo considerando a *demanda não mapeada* como uma *demanda de vertical* adicional.

3.1. Estratégia de planejamento de capacidade

Há diversas estratégias, discutidas na literatura e adotadas na prática, para aquisição de recursos na nuvem. Por exemplo, há estratégias conservadoras, que decidem realizar contratos de compromisso de longo prazo somente depois de usar um determinado tipo de recurso por tempo suficiente para que o contrato seja vantajoso [Wang et al. 2015, Yang et al. 2018]. Há estratégias que tomam como base a demanda realizada no passado para decidir como contratar recursos na nuvem no futuro [den Bossche et al. 2014, Wu et al. 2022]. Em particular, vários clientes de IaaS, inclusive a VTEX, adotam estratégias que usam algumas informações da demanda passada.

No caso da VTEX, a estratégia de aquisição segue um modelo que considera um **período de avaliação** dos contratos. No momento de avaliação, os contratos passados são levados em conta para decidir quais novos contratos serão estabelecidos. Essa decisão é apoiada por dados da demanda passada, considerando o tempo determinado pela **janela de aprendizado**. É analisado qual foi a menor demanda (não nula) por um dado recurso vista durante a **janela de aprendizado**. Esse piso de demanda é considerado um valor firme para a demanda e, portanto, serão adquiridos recursos nos mercados de compromisso de longo prazo para atender essa demanda firme para o futuro. Os contratos futuros têm a duração dada pelo **horizonte de compra**.

Neste estudo, implementando a estratégia de compra da VTEX considerando **período de avaliação**, **janela de aprendizado** e **horizonte de compra** todos com a mesma duração de um ano. Assim, dos três anos avaliados (de 2021 até 2023) usamos o primeiro ano como janela de aprendizado e computamos o custo com os contratos feitos para os dois últimos anos. Em complemento a esta estratégia de compra, consideramos uma segunda estratégia que conhece a demanda futura. Esta estratégia é baseada em um modelo de programação linear que permite escolher de maneira ótima o conjunto de con-

tratos que atendem à demanda com o menor custo possível[1]. Considerar essa segunda estratégia ótima nos permite entender a margem de oportunidade para melhorar a heurística da estratégia adotada pela VTEX. Como não é necessário utilizar a demanda passada na segunda estratégia, consideramos os dois últimos anos (2022 e 2023) na avaliação.

Para ambas as estratégias, utilizamos a demanda de duas formas. Na primeira, calculamos o custo de contratação para a **Demanda global**. Já na segunda forma, usamos as estratégias de planejamento de capacidade aplicadas para calcular a **Demanda por Vertical** para cada uma das 13 verticais de modo independente. Ao fim, calculamos a soma do custo de contratação de cada uma das verticais.

4. Resultados e discussão

Nesta seção, discutimos os resultados do estudo de caso, considerando os custos de aquisição tanto para a estratégia com base em informação do passado (que emula o processo de decisão da VTEX) quanto os custos da estratégia ótima. Apresentamos essa relação em termos do desconto obtido ao adquirir recursos, considerando a demanda global em relação ao custo de aquisição de recursos, considerando as verticais.

Como a Tabela 1 indica, os descontos obtidos ao realizar o planejamento de capacidade considerando a **Demanda Global** são significativos. Primeiro, mesmo usando uma estratégia ótima, ciente da demanda futura, o impacto da variabilidade da demanda das verticais continua relevante: ao agregar a demanda, é possível obter um desconto de quase 7,5%. O impacto da variabilidade na demanda das verticais é ainda mais severo quando consideramos a heurística de alocação usada pela VTEX: ao agrupar a demanda para reduzir essa variabilidade, obtemos um desconto de quase 9% no custo total de aquisição de recursos. Esse maior impacto no custo da heurística acontece justamente por não ser possível antecipar uma mudança de demanda (aumento ou queda) para o ano seguinte, uma vez que o nível de capacidade a comprar nos mercados de longo prazo é baseado em dados do ano anterior. Para além do desconto relativo, a economia de custos é bastante relevante no contexto do estudo de caso: no cenário do estudo de caso, o desconto ao fazer o planejamento de capacidade com a **Demanda Global** equivale a mais de $2,5 milhões.

Tabela 1. Desconto da Aquisição da Demanda Global por Estratégia

Estratégia	Desconto
Heurística	8,86%
Ótima	7,47%

5. Conclusão

Os resultados deste estudo indicam que uma abordagem centralizada para a aquisição de recursos em nuvem em grandes empresas pode gerar grandes economias de custo, tirando melhor proveito da combinação entre os múltiplos mercados oferecidos por provedores IaaS. Considerando uma estratégia de planejamento de capacidade usada na indústria, o desconto potencial, para o caso estudado, foi de quase 9%, o que representa uma economia de milhões de dólares para o caso da demanda estudada. No entanto, como a literatura sugere, uma estrutura descentralizada, que concede alta autonomia às suborganizações,

[1] https://github.com/ufcg-lsd/AWSome-Savings

pode permitir decisões mais rápidas e respostas mais eficazes às mudanças, impactando positivamente o desempenho geral da empresa e até mesmo correlacionando-se com o aumento da velocidade no desenvolvimento de novos produtos. Consideramos expandir esse estudo considerando outras alternativas de combinação da demanda de verticais (além de completamente agregada ou desagregada). É possível que uma combinação mais criteriosa de parte das verticais possa levar a um balanço entre a eficiência da organização e a redução de custos.

6. Agradecimentos

Este trabalho foi financiado pela VTEX BRASIL (EMBRAPII PCEE-2304.0229).

Referências

Carvalho, M., Cirne, W., Brasileiro, F. V., and Wilkes, J. (2014). Long-term slos for reclaimed cloud computing resources. In *Proceedings of the ACM Symposium on Cloud Computing, November 3-5, 2014*, pages 20:1–20:13. ACM.

Costa, R., Brasileiro, F. V., de Souza Filho, G. L., and Sousa, D. M. (2013). Analyzing the impact of elasticity on the profit of cloud computing providers. *Future Gener. Comput. Syst.*, 29(7):1777–1785.

Cullen, J. B. and Perrewé, P. L. (1981). Decision making configurations: An alternative to the centralization/decentralization conceptualization. *Journal of Management*, 7(2):89–103.

den Bossche, R. V., Vanmechelen, K., and Broeckhove, J. (2014). Optimizing iaas reserved contract procurement using load prediction. In *2014 IEEE 7th International Conference on Cloud Computing, June 27 - July 2, 2014*, pages 88–95. IEEE Computer Society.

Englehardt, C. S. and Simmons, P. R. (2002). Organizational flexibility for a changing world. *Leadership & Organization Development Journal*, 23(3):113–121.

Galvão, C., Pereira, T., Brasileiro, F., and Gomes, G. (2024). Costplanner: Planejamento de capacidade de longo prazo para clientes de provedores iaas. In *Anais Estendidos do XLII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 277–282.

Wang, W., Liang, B., and Li, B. (2015). Optimal online multi-instance acquisition in iaas clouds. *IEEE Trans. Parallel Distributed Syst.*, 26(12):3407–3419.

Wu, T., Pan, M., and Yu, Y. (2022). A long-term cloud workload prediction framework for reserved resource allocation. In *IEEE International Conference on Services Computing, SCC 2022, Barcelona, Spain, July 10-16, 2022*, pages 134–139. IEEE.

Yang, S., Pan, L., Wang, Q., Liu, S., and Zhang, S. (2018). Subscription or pay-as-you-go: Optimally purchasing iaas instances in public clouds. In *2018 IEEE ICWS, July 2-7, 2018*, pages 219–226. IEEE.