Planejamento de Capacidade - Definindo SLAs para Sistemas Cloud-Computing

Bernardo Vieira
Instituto de Ciências Exatas e Informática
Sabará, Minas Gerais, Brasil
baavieira@sga.pucminas.br

Gabriel Jota Lizardo
Instituto de Ciências Exatas e Informática
Nova Lima, Minas Gerais, Brasil
gabriel.jota@sga.pucminas.br

Guilherme Oliveira de Rodrigues
Instituto de Ciências Exatas e Informática
Nova Lima, Minas Gerais, Brasil
guilherme.rodrigues.1449815@sga.pucminas.br

Henrique Oliveira da C. Franco
Instituto de Ciências Exatas e Informática
Belo Horizonte, Minas Gerais, Brasil
henrique.franco@sga.pucminas.br

Larissa Mariella
Instituto de Ciências Exatas e Informática
Matozinhos, Minas Gerais, Brasil
larissa.mariella@sga.pucminas.br

Rondinelly Martins
Instituto de Ciências Exatas e Informática
Belo Horizonte, Minas Gerais, Brasil
1458176@sga.pucminas.br

ABSTRACT

This report presents a strategic framework for Bingus, a pet product company, to adopt cloud services for managing its expanding digital operations. Bingus, which specializes in smart toys, organic foods, and sustainable accessories for pets, seeks to use cloud technology to handle data on customers, inventory, and logistics. Key recommendations for Service Level Agreements (SLAs) are provided to enhance user experience and operational reliability.

Suggested SLAs focus on response time (1-2 seconds), throughput (100-500 tps), and availability (99.9%), balancing user satisfaction and cost control. Cost estimates using exponential and linear models show ways to manage expenses under varying service demands.

To optimize costs, the report recommends negotiating SLAs with lower thresholds (e.g., 3-5 seconds response, 99% availability) when feasible, and employing auto-scaling and SLA adjustments for better cost efficiency. A priority framework emphasizes response time and availability, ensuring Bingus delivers a dependable cloud service within budget.

KEYWORDS

cloud computing, service level agreement, análise de dados

ACM Reference Format:
Bernardo Vieira, Gabriel Jota Lizardo, Guilherme Oliveira de Rodrigues, Henrique Oliveira da C. Franco, Larissa Mariella, and Rondinelly Martins. 2024. Planejamento de Capacidade - Definindo SLAs para Sistemas Cloud-Computing. In *Proceedings of ACM Conference (.).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

., November 2024, Belo Horizonte

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

1 SLAS SUGERIDOS

- **SLAr (Tempo médio de resposta por transação)**:
  - **Objetivo**: Manter o tempo de resposta de cada transação de compra ou navegação no site dentro de uma média aceitável para garantir que os usuários não enfrentem lentidão.
  - **Valor sugerido**: 1 a 2 segundos por transação. Esse intervalo é crucial para evitar abandonos de carrinho e manter a experiência de navegação eficiente.
- **SLAx (Throughput da transação em tps - transações por segundo)**:
  - **Objetivo**: Garantir que o sistema suporte um grande número de transações simultâneas, especialmente em horários de picos, como promoções e datas comemorativas.
  - **Valor sugerido**: 100 a 200 tps em condições normais, com possibilidade de escalar até 500 tps durante eventos promocionais. Essa capacidade assegura que o sistema possa lidar com altas demandas sem interrupções.
- **SLAa (Disponibilidade na nuvem)**:
  - **Objetivo**: Maximizar a disponibilidade do serviço para que os usuários possam acessar o site e outros serviços da Bingus a qualquer momento, independentemente da localização.
  - **Valor sugerido**: 99,9% ou mais. Esse valor garante que as interrupções sejam mínimas e que a Bingus mantenha uma presença constante e confiável, essencial para o e-commerce e serviços que requerem disponibilidade contínua.

Esses SLAs são projetados para garantir um ambiente robusto e de alta qualidade, alinhado com as expectativas dos clientes modernos e as melhores práticas de mercado para serviços em *cloud*.

2 CÁLCULO DE CUSTOS

Com base nas fórmulas e tabelas fornecidas, os custos por requisição no mercado de *cloud* podem ser estimados da seguinte forma:

(1) **Custo por transação do tempo de resposta ($C_r(\mathbf{SLA}_r)$)**:

$$C_r(\text{SLA}_r) = \alpha_r e^{-\beta_r \text{SLA}_r}$$

indica que o custo diminui exponencialmente à medida que o SLA_r (tempo médio de resposta) aumenta.

(2) **Custo por transação do throughput ($C_x(\mathbf{SLA}_x)$)**:

$$C_x(\text{SLA}_x) = \alpha_x \text{SLA}_x$$

mostra que o custo é proporcional ao throughput negociado.

(3) **Custo por transação de disponibilidade ($C_a(\mathbf{SLA}_a)$)**:

$$C_a(\text{SLA}_a) = e^{\beta_a \text{SLA}_a} - e^{0.9\beta_a}, \quad \text{SLA}_a \geq 0.9$$

Os custos são mais elevados conforme a disponibilidade desejada aumenta.

3 ESTIMATIVAS DE MERCADO

- **Tempo de resposta (SLAr)**: Custos podem variar de forma significativa dependendo da exigência do tempo de resposta. Custos baixos (próximo de 1 centavo por transação) são aplicáveis a tempos de resposta menos exigentes (2 a 5 segundos), enquanto tempos de resposta mais rigorosos têm custos maiores.
- **Throughput (SLAx)**: Custos são lineares, com variações entre 1 e 5 centavos por transação, dependendo do nível de throughput necessário. Por exemplo, 100 transações por segundo podem ter um custo de 100 a 500 centavos, dependendo da escolha do coeficiente.
- **Disponibilidade (SLAa)**: Custos aumentam exponencialmente conforme o SLA se aproxima de 1 (100% de disponibilidade), refletindo a complexidade e o custo elevado de garantir disponibilidade máxima.

4 PLANEJAMENTO DE SLAS COM RECURSOS FINANCEIROS RESTRITOS

Para otimizar o planejamento de SLAs e equilibrar desempenho e custos, seguem as recomendações:

- **Tempo de resposta (SLAr)**:
  - **Opção**: Negociar SLAs de tempo de resposta que mantenham um nível aceitável de performance, como 3 a 5 segundos, com custos menores.
  - **Justificativa**: Redução de custos por transação em até 60-80%, dependendo de α_r e β_r.
- **Throughput (SLAx)**:
  - **Opção**: Manter um throughput médio de 100 a 200 tps com escalabilidade automática para picos.
  - **Justificativa**: Controle de custos, com a possibilidade de escalar conforme necessário.
- **Disponibilidade (SLAa)**:
  - **Opção**: Negociar uma disponibilidade de 99% em vez de 99,9% ou 100%.
  - **Justificativa**: Redução significativa de custos, mantendo alta confiabilidade.

5 IMPLEMENTAÇÃO TEÓRICA

Levando em conta o tipo de negócio da empresa em questão, a implementação realizada pelo nosso grupo abordou os seguintes valores:

Tabela de Dados dos SLAs e Pesos

SLA	Valores	Pesos
Tempo de Resposta (SLAr)	1	0,2
Taxa de Processamento (SLAx)	116,156	0,2
Disponibilidade (SLAa)	0,99	0,6

Table 1: Dados dos SLAs e seus respectivos valores e pesos.

Tabela de Função de Utilidade e Custo Total

Métrica	Valor
Função de Utilidade	0,84757
Custo Total	4

Table 2: Função de utilidade e custo total.

Tabela de Coeficientes

Coeficiente	Valores
α_r	0,4
β_r	0,1
α_x	0,03
β_a	0,8

Table 3: Valores dos coeficientes utilizados nas fórmulas de custo.

w_r	w_x	w_a	Util
0,2	0,2	0,6	0,84757
0,3	0,3	0,4	0,82136
0,4	0,4	0,2	0,79515
0,5	0,45	0,05	0,76394
0,6	0,3	0,1	0,71273
0,01	0,01	0,98	0,89738

Table 4: Tabela de pesos dos SLAs e valores da função de utilidade.

6 GRÁFICOS DOS CUSTOS

A seguir, estão alguns gráficos gerados no Excel por meio de ferramentes de plotagem gráfica.

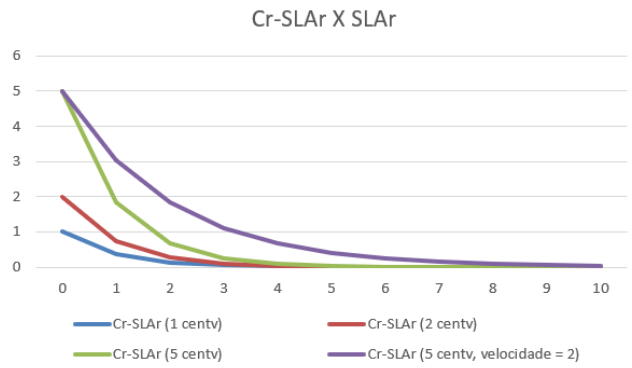

Figure 1: Gráfico de Custo x SlaR

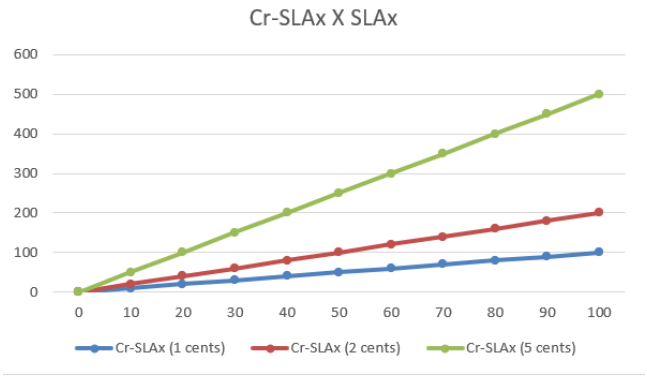

Figure 2: Gráfico de CustoX x SlaX

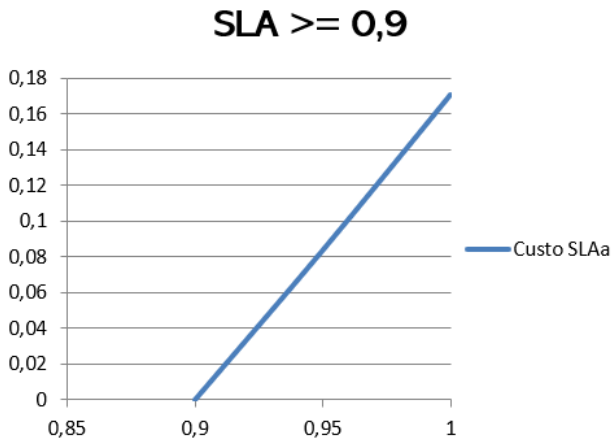

Figure 3: Gráfico de CustoA x SlaA

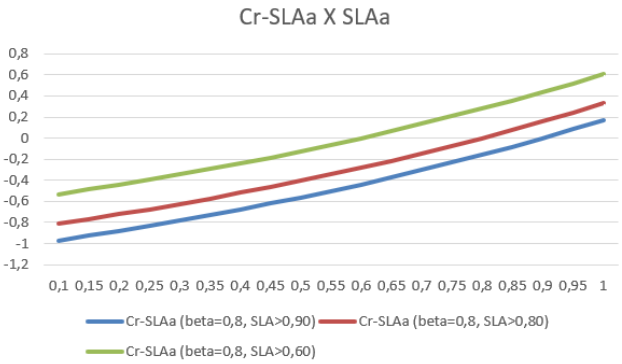

Figure 4: Gráfico de CustoA x SlaA com múltiplos SLA's

Link da planilha Excel: Link

Link para download no Drive: Link Drive