Planejamento de Capacidade: Definindo SLAs para Infraestrutura em Nuvem

---

Alessandra Faria Rodrigues
(828333)

Novembro de 2025

1. **Proposta da Empresa Fictícia**

Empresa: FinTech Smart
Contexto: A "FinTech Smart" é uma startup B2B que oferece uma API de análise de risco de crédito em tempo real. Nossos clientes (outras fintechs, bancos digitais e plataformas de e-commerce) integram nossa API em seus checkouts para aprovar ou negar transações financeiras instantaneamente.
Necessidade de Cloud: O serviço é crítico e possui demanda flutuante (picos durante eventos de vendas como Black Friday e vales em madrugadas). A computação em nuvem foi escolhida por sua escalabilidade, modelo de custo "pay-as-you-go" e alta disponibilidade gerenciada, eliminando a necessidade de grandes investimentos iniciais em hardware.
O Desafio: Precisamos definir o Acordo de Nível de Serviço (SLA) ideal com nosso provedor de nuvem. Devemos equilibrar três métricas-chave dentro de um **orçamento restrito de R$ 20.000,00 por mês**:

1. **SLAr:** Tempo de Resposta (em segundos).
2. **SLAx:** Taxa de Processamento (em requisições/segundo).
3. **SLAa:** Disponibilidade (em percentual).

O objetivo é maximizar a **Qualidade (Utilidade)** do nosso serviço para os clientes, otimizando a relação custo-benefício.

2. **Definição e Justificativa dos SLAs**

- **Tempo de Resposta (SLAr): Alta Prioridade.** Uma resposta lenta da nossa API significa que o cliente final (o comprador no e-commerce) ficará esperando pela aprovação da transação. Tempos acima de 1-2 segundos são considerados inaceitáveis e levam ao abandono de carrinho.
- **Disponibilidade (SLAa): Alta Prioridade.** Se nossa API estiver indisponível, nenhuma transação pode ser processada. Isso causa interrupção direta nos negócios dos nossos clientes.
- **Taxa de Processamento (SLAx): Média/Alta Prioridade.** Devemos ser capazes de lidar com o volume de requisições, especialmente em horários de pico. Uma taxa de processamento inadequada gera filas e, por consequência, aumenta o tempo de resposta.

Estratégia de Priorização Escolhida: Em vez de priorizar um único fator, optamos por uma abordagem mais equilibrada. Nós percebemos a qualidade (Utilidade) como uma combinação dos três fatores. Portanto, atribuímos pesos iguais ($wr = wx = wa = 0.333\ldots$) para cada SLA na nossa função de otimização. O objetivo é encontrar o ponto de equilíbrio ideal entre os três.

3. Estimativa de Custos de Mercado e Metodologia

Utilizando os dados de custos fornecidos pelo provedor (baseados nos arquivos Custos.csv), que demonstram uma relação não-linear entre custo e performance:

- **Custo(SLAr):** O custo aumenta exponencialmente quanto menor o tempo de resposta exigido (chegar a 0.1s é muito mais caro que 1s).
- **Custo(SLAx):** O custo aumenta quanto maior a taxa de processamento garantida.
- **Custo(SLAa):** O custo aumenta exponencialmente à medida que a disponibilidade se aproxima de 100% (passar de 99% para 99.9% é significativamente mais caro).

Função de Otimização (Utilidade): Para simular a "qualidade" do serviço, utilizamos a função de utilidade (do arquivo Utiliti.csv):

$$U = w_r \cdot \left(\frac{2.0 \cdot e^{-SLAr}}{1+e^{-SLAr}}\right) + w_x \cdot \left(1 - e^{-0.1 \cdot SLAx}\right) + w_a \cdot (10 \cdot SLAa - 9)$$

Onde $(wr = wx = wa = 0.333\ldots)$ (pesos iguais).

Otimização: Utilizamos uma ferramenta de otimização (Solver) para encontrar os valores de SLAr, SLAx e SLAa que **maximizam a Função Utilidade**, sujeitos à minimização do Custo Total por requisição.

4. Resultados da Simulação e Otimização

A simulação (baseada no arquivo Solver.csv) encontrou o seguinte ponto ótimo para a nossa configuração balanceada:

Métrica de SLA	Valor Ótimo	Justificativa
SLAr (Tempo de Resposta)	1 segundo	Oferece uma excelente experiência de usuário para transações financeiras sem incorrer nos custos exponenciais de tempos de resposta sub-segundo.
SLAx (Taxa de Processamento)	50 req/segundo	Um valor robusto que cobre a demanda média e picos moderados da nossa base de clientes B2B.
SLAa (Disponibilidade)	0.99 (99,0%)	Garante que o serviço esteja operacional na vasta maioria do tempo.

Tabela 1: Resultados de otimização obtidos via Solver

Métricas de Resultado (do Solver):

- Qualidade (Utilidade) Máxima: \$0.8104\$ (score de utilidade máximo alcançado com esta configuração).
- Custo Total por Requisição: \$2.0153\$ centavos (ou R$ 0,020153)

5. **Análise de Orçamento e Conclusão**

O principal fator limitante é o nosso orçamento mensal de R$ 20.000,00. Com base no custo por requisição otimizado, podemos calcular nossa capacidade máxima de processamento:

- **Custo por Requisição:** R$ 0,020153
- **Orçamento Mensal:** R$ 20.000,00

Capacidade Máxima = Orçamento / Custo por Requisição
Capacidade Máxima = R$ 20.000,00 / R$ 0,020153 ~ **992.408 requisições por mês**

Recomendação: Recomenda-se a contratação do serviço em nuvem com os seguintes SLAs:

- Tempo de Resposta (SLAr): 1 segundo
- Taxa de Processamento (SLAx): 50 req/seg
- Disponibilidade (SLAa): 99%

Esta configuração oferece a melhor qualidade (Utilidade = 0.8104) para nossos clientes, dado nosso modelo de negócio "balanceado", e nos permite processar aproximadamente **992.400 transações por mês** dentro do orçamento estipulado de R$ 20.000,00.