Confiabilidade para crescer sem perder controle

Uma evolução pragmática de observabilidade, SLOs e incident management para um time SRE inicial

Luis Cruz - Analista de Infraestrutura e Segurança | DevOps & SRE

BASE	FOCO	RITMO
Prometheus + Grafana + Loki	**Dois serviços piloto**	**30 / 60 / 90 dias**

Três decisões criam o sistema operacional de confiabilidade.

01

Preservar

Prometheus, Grafana e Loki continuam como base. O problema não é falta de ferramenta.

02

Governar

SLOs, error budget e burn rate conectam operação ao impacto real.

03

Evoluir

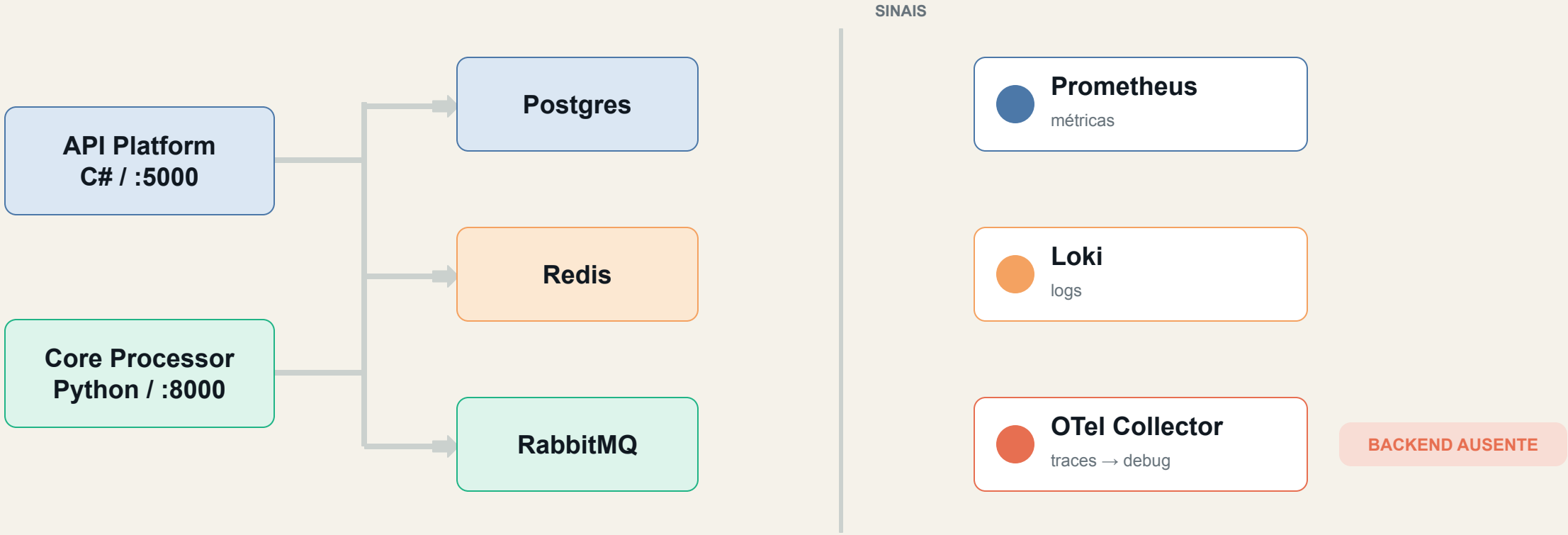
OpenTelemetry recebe backend e sampling antes de qualquer expansão ampla.

Resultado esperado

Menor MTTD e MTTR, menos alertas sem ação e adoção de tracing com custo previsível.

ESTADO ATUAL

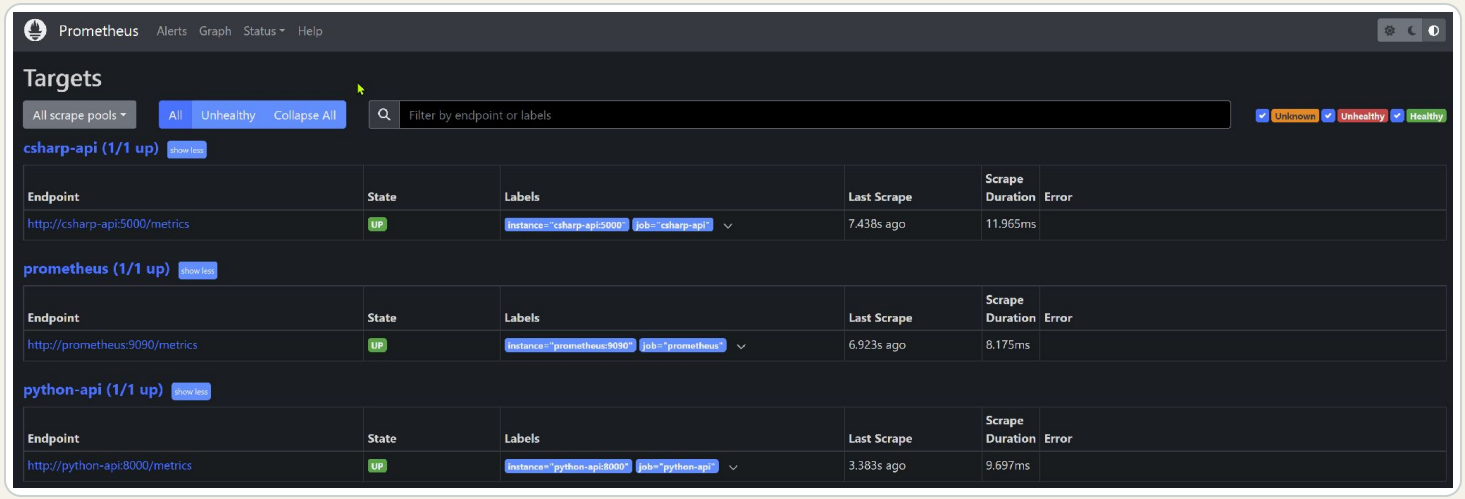
A stack cobre aplicações e dependências, mas os sinais ainda não formam uma jornada única.

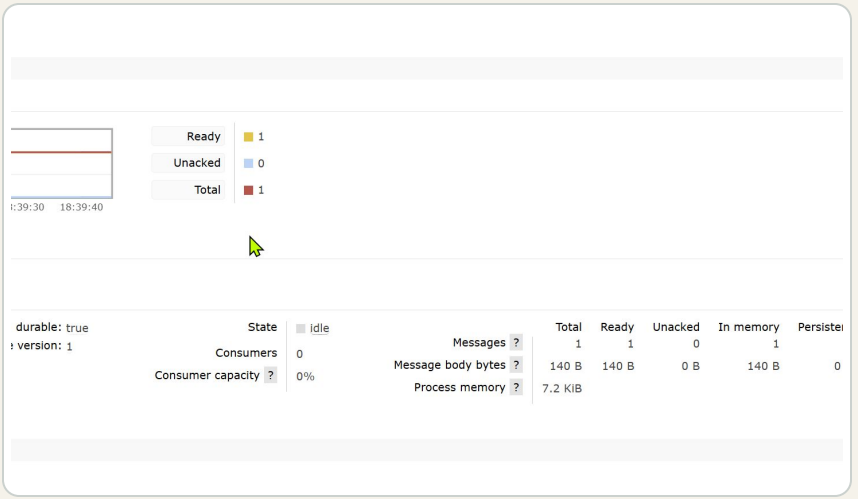


docker-compose.yml; apps/*; monitoring/*

EVIDÊNCIA

A base funciona localmente; as lacunas observadas são operacionais.





3/3

targets UP

coleta validada

2

APIs com logs no Loki

coleta validada

1

mensagem Ready

fila observada

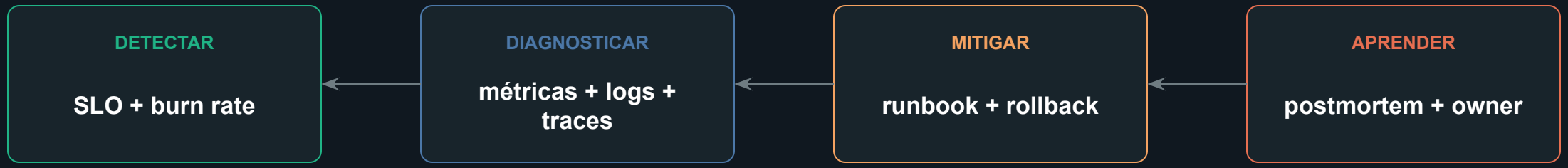
0

consumers ativos

fila observada

GAP DE MATURIDADE

Coletar sinais não basta: o ciclo precisa terminar em aprendizado.



Hoje

Métricas e logs existem; tracing, alertas por impacto e governança de ações ainda estão incompletos.

PRINCÍPIO

Melhorar a qualidade dos sinais vem antes de aumentar a quantidade de ferramentas.

1

QUALIDADE

Padronizar labels, request_id, service.name e dashboards pelos quatro sinais.

2

GOVERNANÇA

Definir SLOs, alertas acionáveis, runbooks e donos para ações.

3

CORRELAÇÃO

Fechar métricas → logs → traces nos fluxos críticos.

4

ESCALA

Expandir retenção, serviços e APM somente após medir ROI.

SLO PILOTO

API Platform: experiência do usuário vira critério de deploy.

99,5%

Disponibilidade

≈ 3h36 de budget em 30 dias

< 500ms

Latência P95

janelas móveis de 5 minutos

< 1%

Erros 5xx

percentual mensal

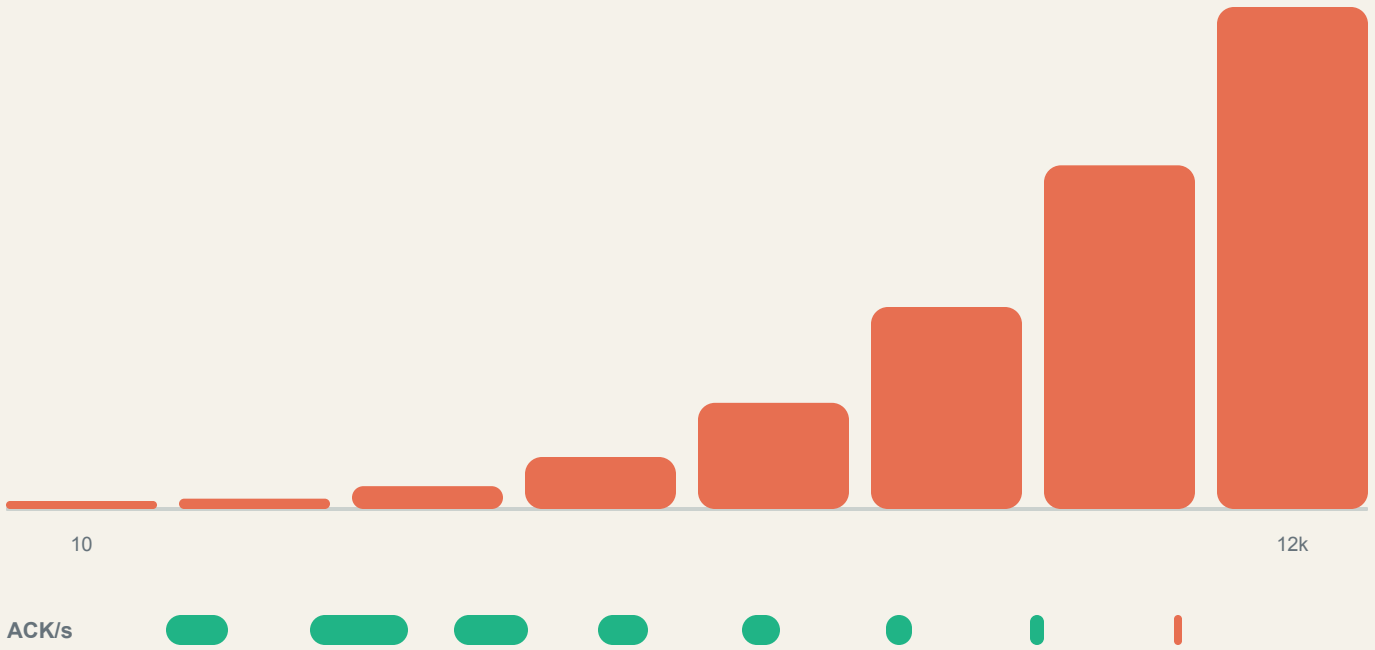
Como o SLO muda a operação

- **Budget saudável**
  deploy normal
- **Consumo acelerado**
  reduzir risco
- **Budget esgotado**
  priorizar confiabilidade

SLO PILOTO

Core Processor: estar de pé não significa que o trabalho está progredindo.

BACKLOG DA FILA



< 1.000

mensagens por >10 min

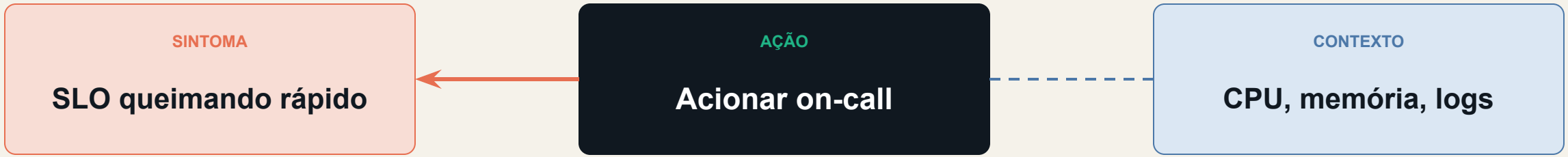
< 5 min

P95 da idade

≥ 99%

ACK com sucesso

A página deve representar impacto; os sinais técnicos explicam o porquê.

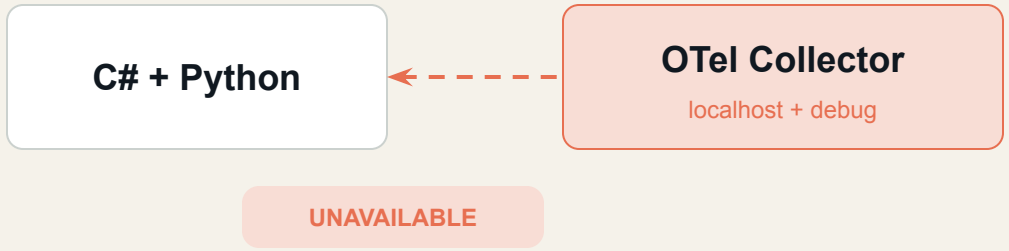


MULTI-WINDOW

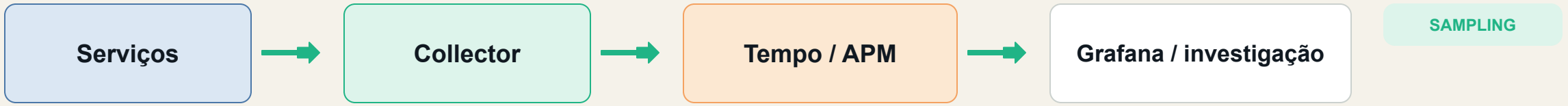
5m + 1h	página	
30m + 6h		investigar
24h		tendência

OpenTelemetry é a abstração certa; falta fechar o transporte e a consulta.

HOJE



ALVO



DECISÃO DE APM

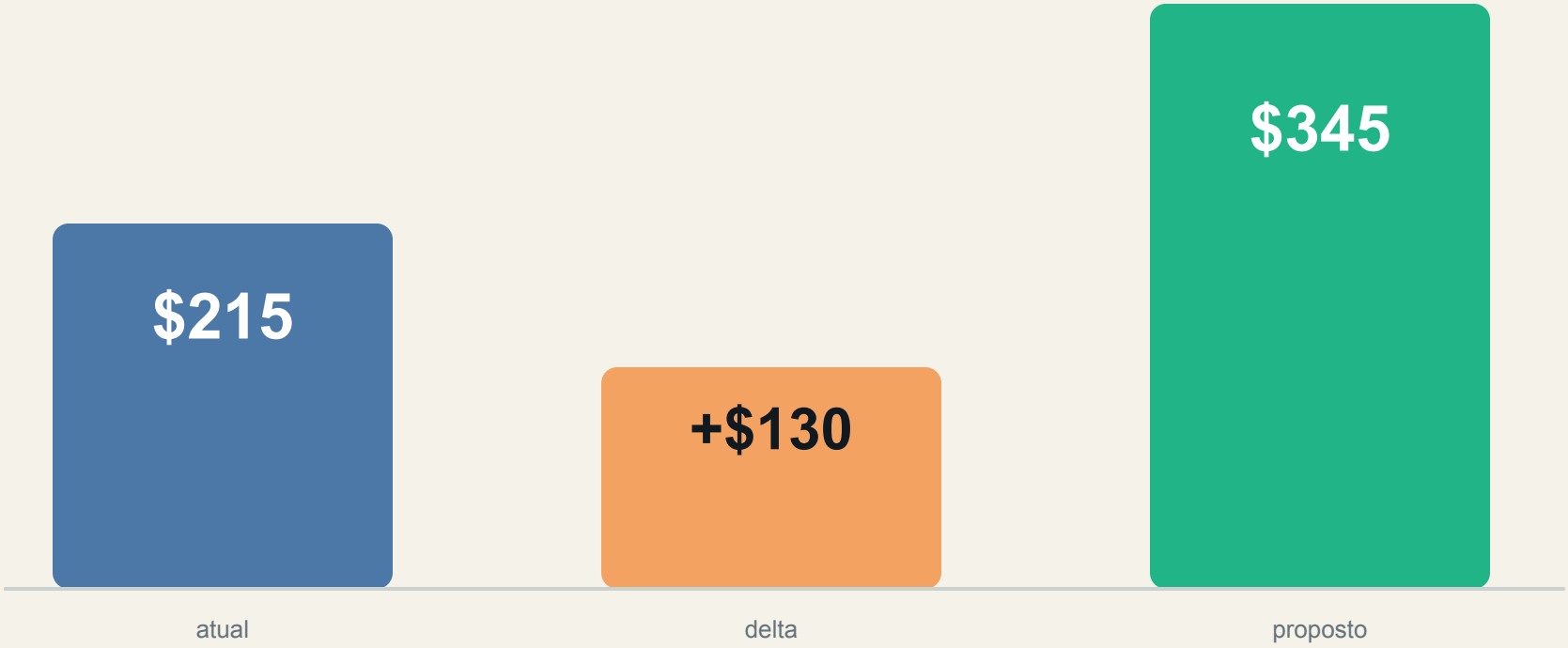
Tempo/Grafana é o piloto pragmático; vendor comercial é uma opção de aceleração.

Opção	Adoção	Custo	Lock-in	Operação
Tempo + OTel	média	baixo/médio	baixo	time assume
Datadog	rápida	alto	médio/alto	gerenciada
New Relic	rápida	médio/alto	médio	gerenciada
Dynatrace	média	alto	alto	enterprise

Decisão condicionada — **Expandir somente após medir MTTR, adoção e custo por GB útil.**

FINOPS

O piloto adiciona ~US$130/mês; tracing concentra o novo custo.



3 alavancas

Sampling

100% erros / ~10% normal

Retenção

7 dias para logs brutos

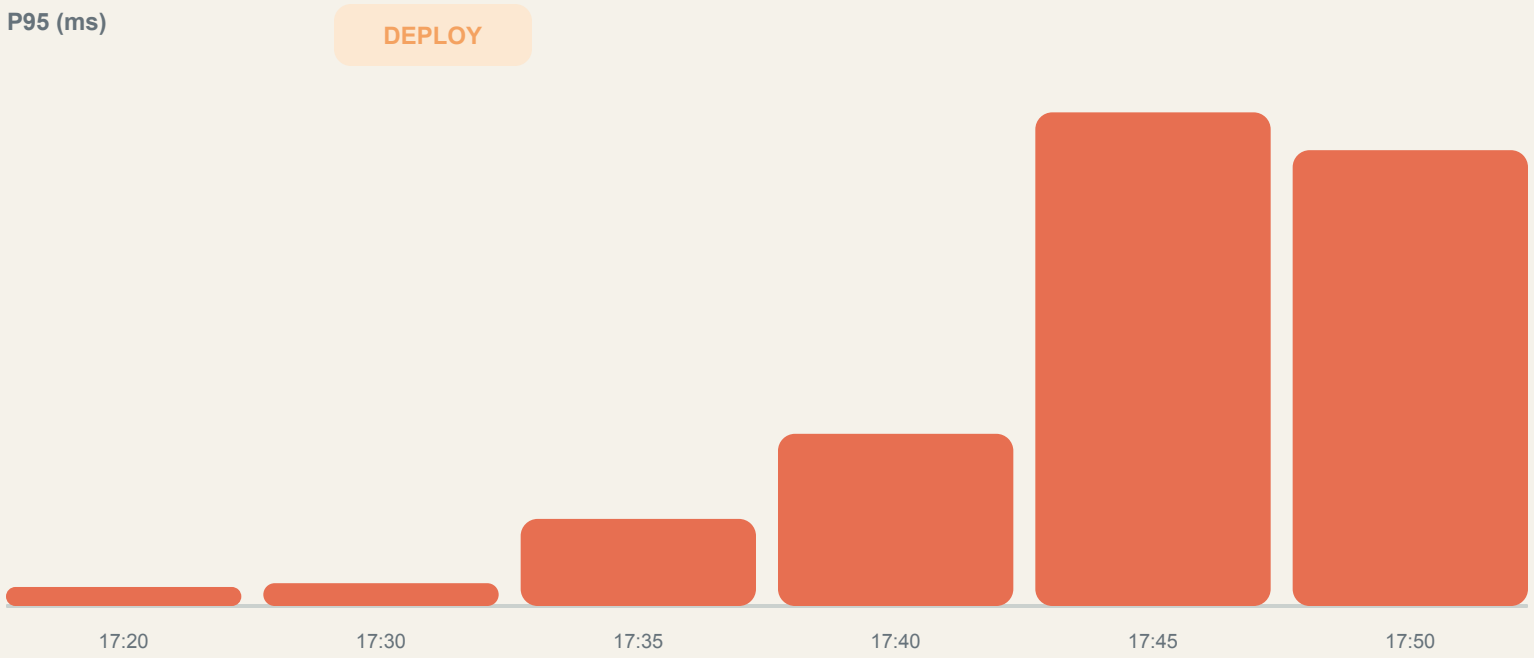
Cardinalidade

sem IDs em labels

Estimativa para piloto; calibrar com telemetria real e cotação comercial.

INCIDENTE

Após o deploy, latência, erros e queries degradam na mesma janela.



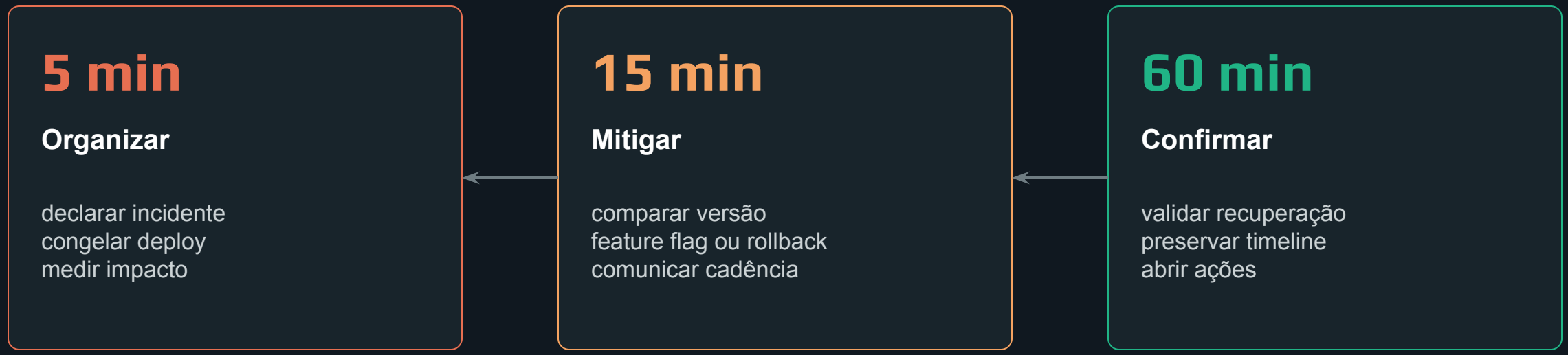
5,2s
P95 máximo
17:45

9,0s
P99 máximo
17:45

6,0%
erros
17:45

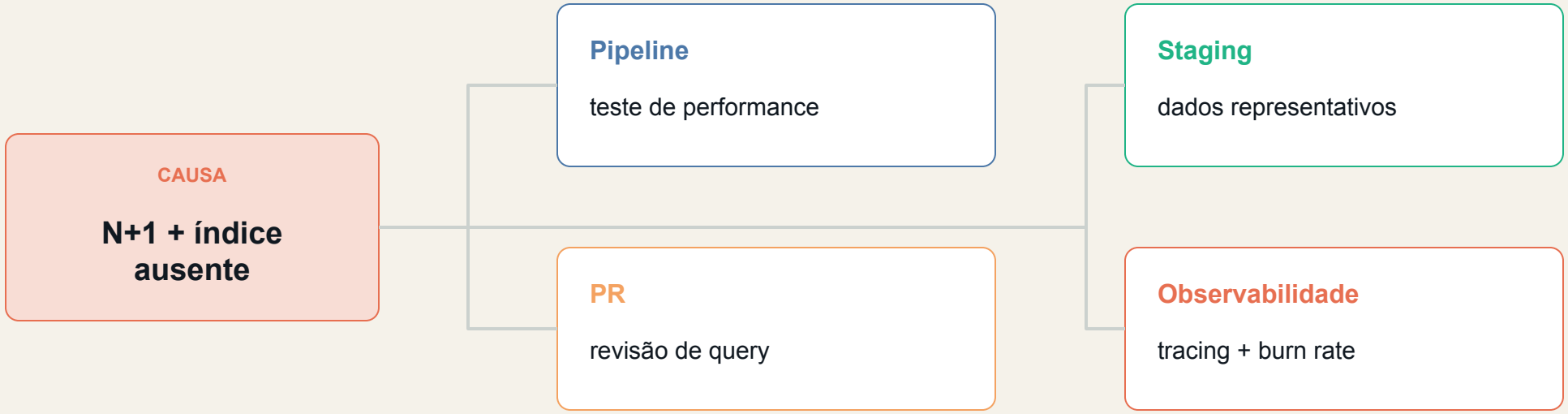
5,1s
slow query
users.account_id

O ritmo 5 / 15 / 60 restaura o serviço antes de perseguir certeza.



Rollback é controle de dano, não falha do time.

O N+1 é a causa técnica; a prevenção precisa corrigir o sistema de entrega.



Blameless Donos e prazos transformam aprendizado em redução de recorrência.

PRIMEIROS 90 DIAS

O primeiro ciclo entrega um mínimo sistema operacional de confiabilidade.

0–30 dias

Estabilizar

- dashboards mínimos
- labels e cardinalidade
- 3 runbooks

31–60 dias

Governar

- 2 SLOs piloto
- burn-rate alerts
- ritual de incidentes

61–90 dias

Correlacionar

- backend de traces
- sampling
- ROI do piloto

Gate de saída

Sinais confiáveis, on-call acionável e diagnóstico do caminho crítico em uma única jornada.

ESCALA DE MATURIDADE

A expansão acontece depois que o piloto prova adoção, confiabilidade e ROI.

3 meses	6 meses	12 meses	18 meses
Piloto	**Governança**	**Expansão**	**Otimização**
2 serviços SLO + tracing	canary error budget	mais jornadas self-service	automação FinOps contínuo
1	2	3	4

A função SRE não vira um gargalo central: padrões e ownership permanecem com os times de serviço.

FECHAMENTO

Confiabilidade primeiro. Complexidade depois.

1. Preservar a base que já funciona.
2. Medir experiência real com SLOs.
3. Expandir tooling somente com ROI.