THREE HORIZONS, O CAMINHO RED HAT · PLAYBOOK

Do Azure cru a uma Agentic DevOps Platform em 90 a 180 dias.

O Three Horizons é o acelerador para empresas que querem a plataforma com suporte de fornecedor, certificada e com propriedade conjunta de Red Hat e Microsoft: Red Hat Developer Hub no Azure Red Hat OpenShift, OpenShift Pipelines e GitOps, Trusted Application Pipeline, e a fundação de IA da Microsoft por baixo. Este playbook percorre a stack opinada, a sequência de três horizontes da fundação segura aos agentes governados, e o framework de decisão que diz quando este caminho é o certo.

AUTORA

Paula Silva

CARGO

AI-Native Software Engineer

CONTATO

in/paulanunes

TEMPO DE LEITURA

~ 40 minutos, fim a fim

3

HORIZONTES

6

CAPÍTULOS

22

TEMPLATES DE GOLDEN PATH

90-180

DIAS ATÉ O PRIMEIRO VALOR EM PRODUÇÃO

CAPÍTULOS

Chapter 00

O argumento

O que é o acelerador, a proposta de valor em três partes (velocidade, coerência comercial, postura regulatória), os números de referência e a escada de maturidade que o batiza.

O Argumento →

Chapter 01

A stack opinada

RHDH como IDP, ARO como cluster operado conjuntamente, Pipelines mais GitOps como entrega, Trusted Application Pipeline como cadeia de suprimentos, e a camada de assistência de IA em três superfícies.

Arquitetura →

Chapter 02

Horizonte 1, a fundação

O runtime Azure seguro em 90 dias: rede zero-trust, nenhuma credencial estática em lugar nenhum, registry e dados em private endpoints, recuperação testada e governança de custos desde o primeiro dia.

Horizonte 1 →

Chapter 03

Horizonte 2, o produto interno

GitOps app-of-apps, Developer Hub como porta de entrada, o catálogo de 22 templates de Golden Path, e a stack de observabilidade com mais de 30 alertas em cinco camadas.

Horizonte 2 →

Chapter 04

Horizonte 3, a camada agêntica

Azure AI Foundry como Terraform, sete Golden Paths de IA, a frota de 17 agentes com integrações MCP, e os gates de avaliação que tornam a autonomia governável.

Horizonte 3 →

Chapter 05

A decisão e a implantação

O crosswalk dos Quatro Pilares, a estrutura comercial conjunta, os perfis de cliente que pendem para cada lado, e a sequência de três fases ancorada no marco de primeiro valor.

A Decisão →

DICAS DE LEITURA

ATALHOS DE TECLADO

Aperte `/` para buscar. Use `j` e `k` para mover entre capítulos. Aperte `t` para alternar tema. Aperte `g` para ir ao topo.

Paula Silva

AI-NATIVE SOFTWARE ENGINEER

in/paulanunes

v1.0.0 · 2026-07-03

Building the future of software development with AI and Agentic DevOps

00 O ARGUMENTO

Um acelerador, uma stack opinada: do Azure cru a uma Agentic DevOps Platform.

O Three Horizons é um acelerador implantado pelo cliente que materializa uma Agentic DevOps Platform na stack Red Hat rodando no Azure. Ele existe porque montar essa plataforma a partir de componentes crus leva de 9 a 18 meses, e porque uma classe de empresas precisa que o resultado seja suportado pelo fornecedor, certificado e auditável desde o primeiro dia. Este capítulo define o que é o acelerador, enuncia sua proposta de valor e apresenta a escada de maturidade que lhe dá nome.

O que é o Three Horizons

O Three Horizons é uma arquitetura de referência e um padrão de implantação pelo cliente, não um produto com SKU. Ele combina produtos existentes de Red Hat e Microsoft em uma fundação integrada e opinada: Red Hat Developer Hub como Internal Developer Platform, Azure Red Hat OpenShift como cluster, OpenShift Pipelines e OpenShift GitOps como substrato de entrega, Red Hat Trusted Application Pipeline para segurança de cadeia de suprimentos, e a fundação Microsoft (Entra ID, Azure AI Foundry, GitHub Enterprise) por baixo. Os componentes são licenciados e suportados de forma independente; o acelerador é a composição.

Fisicamente, o acelerador é entregue como um único repositório template no GitHub: mais de 120 arquivos e cerca de 20 mil linhas de código curado e documentado, incluindo 16 módulos Terraform que cobrem todas as camadas de infraestrutura Azure, 22 templates de Golden Path para scaffolding self-service, 17 agentes especializados de Copilot Chat, 13 integrações de servidores MCP e mais de 30 regras de alerta Prometheus. O cliente clona o template e customiza política e conteúdo. O cliente não monta integrações.

> DEFINIÇÃO
>
> O princípio operacional de todo acelerador deste corpo de trabalho: o cliente customiza, o cliente não monta. O trabalho de integração é a parte mais cara da construção de plataforma, e é exatamente a parte que o acelerador absorve.

A proposta de valor em três partes

O primeiro componente é velocidade. Uma plataforma green-field que levaria de 9 a 18 meses para ser montada a partir de componentes crus comprime para 90 a 180 dias, porque o acelerador entrega os padrões de integração, o catálogo de Golden Paths, a fiação de cadeia de suprimentos, a configuração GitOps e a infraestrutura de agentes pré-integrados. O segundo componente é coerência comercial. Red Hat e Microsoft operam os serviços subjacentes como parceria estratégica conjunta, então o cliente tem uma implantação de fundação com dois donos comerciais de classe mundial e conversas unificadas de suporte, licenciamento e roadmap.

O terceiro componente é postura regulatória e de auditoria. A stack Red Hat carrega certificações FedRAMP, FIPS e Common Criteria que alguns setores tratam como critério bloqueante de compras, e o acelerador entrega defaults de política opinados para regimes latino-americanos: LGPD no Brasil, BACEN em serviços financeiros, ANEEL em energia, ANS em saúde. O cliente herda essa postura em vez de negociá-la. Para bancos, utilities, saúde e governo, esse terceiro componente é frequentemente o decisivo.

O impacto, em números

As implantações de referência do acelerador relatam quatro efeitos principais. Velocidade de deploy: releases cerca de 60% mais rápidos, porque pipelines Tekton automatizados e promoção via GitOps cortam ciclos de semanas para horas. Eficiência operacional: cerca de 40% menos toil, porque infraestrutura como código, reconciliação e portais self-service removem trabalho manual da semana do time de plataforma. Experiência do desenvolvedor: pesquisas de adoção relatam ganhos de satisfação em torno de 3x quando Golden Paths e fluxos assistidos por Copilot chegam. Compliance: política como código coloca todo deploy sob os mesmos padrões aplicados.

Trate esses números como um líder de engenharia deve tratar: evidência direcional de implantações de referência, não garantias. Os compromissos mensuráveis que importam são os que a própria plataforma emite quando está rodando: lead time e taxa de falha de mudança DORA para os times onboardados, atribuição de custo por workload nos dashboards de gasto, e o marco de primeiro valor, a semana em que o primeiro time entrega por um Golden Path mais rápido que seu próprio baseline anterior.

> DEFINIÇÃO
>
> Quatro números de referência: 60% mais velocidade de deploy, 40% menos toil operacional, 3x de satisfação do desenvolvedor e 100% dos deploys sob política como código. Verifique-os no seu próprio parque com métricas DORA e atribuição de custo, ambas entregues por padrão pela plataforma.

A escada de maturidade que dá nome ao acelerador

O Three Horizons é batizado pelo modelo de maturidade que operacionaliza. Horizonte 1, Otimizar o Presente, é a fundação crível: cluster, rede, identidade, pipelines e a linha de base de segurança, mirando os primeiros 90 dias. Horizonte 2, Capacidades Emergentes, transforma infraestrutura em produto interno: o portal Developer Hub, Golden Paths em escala, observabilidade e self-service governado, nos meses três a seis. Horizonte 3, Transformar o Futuro, é a camada agêntica: AI Foundry, workloads RAG e multiagente, e operações autônomas com revisão humana nas fronteiras de promoção, nos meses seis a doze.

A escada é um modelo de sequenciamento: diz ao time de plataforma o que construir a seguir e se recusa a deixar a camada agêntica pousar sobre uma fundação inacabada. Ela é compatível com o CNCF Platform Maturity Model, que é um modelo de medição: o Horizonte 1 corresponde aproximadamente ao nível Operational, o Horizonte 2 ao Scalable e o Horizonte 3 ao Optimizing ou AI-Native na maioria das dimensões. Os capítulos 02 a 04 percorrem cada horizonte em profundidade operacional, e o capítulo 05 fecha com o framework de decisão e a sequência de implantação.

Referências

Fontes primárias para a definição do acelerador, a stack e o modelo de maturidade.

- Documentação do Red Hat Developer Hub, a distribuição enterprise do Backstage que serve de IDP ao acelerador.
- Documentação do Azure Red Hat OpenShift, o serviço OpenShift gerenciado e operado conjuntamente em que a plataforma roda.
- CNCF Platforms Whitepaper, a definição de referência das capacidades de plataforma que o acelerador materializa.
- CNCF Platform Engineering Maturity Model, o modelo de medição sobre o qual a escada H1/H2/H3 se mapeia.
- DORA 2025, State of AI-assisted Software Development, a lente de métricas de entrega usada para verificar as alegações de impacto do acelerador.
- Red Hat Trusted Application Pipeline, a stack de segurança de cadeia de suprimentos ligada por padrão.

Paula Silva

AI-NATIVE SOFTWARE ENGINEER

in/paulanunes

v1.0.0 · 2026-07-03

Building the future of software development with AI and Agentic DevOps

01 ARQUITETURA

Cada componente nomeado, cada componente integrado: a stack de relance.

A stack do Three Horizons é opinada de propósito. Cada camada tem um componente nomeado, e cada componente chega integrado aos demais no momento do deploy. Este capítulo percorre as cinco escolhas estruturais: a Internal Developer Platform, o cluster, o substrato de entrega, a stack de segurança de cadeia de suprimentos e a camada de assistência de IA, e explica o que cada escolha compra para o cliente.

Red Hat Developer Hub como IDP

O Red Hat Developer Hub é a distribuição enterprise do Backstage suportada pela Red Hat, e faz para o Backstage upstream o mesmo que o Red Hat Enterprise Linux faz para o Linux upstream: mesmo núcleo, empacotamento endurecido, suporte enterprise, ciclo de vida mais longo. O RHDH entrega um conjunto curado de cerca de 30 a 40 plugins testados, incluindo plugins nativos de OpenShift que trazem informações de Pipelines, GitOps, ACM e ACS direto para o catálogo, além de suporte 24x7 com patching de segurança e assistência de upgrade. Para indústrias reguladas, esse contrato de suporte é critério de compras, não conveniência.

Os Software Templates do RHDH são a superfície de Golden Paths do acelerador, organizados pela escada H1/H2/H3, e o Software Catalog é o registro de cada serviço, componente, API e recurso da empresa, populado por descoberta automática e por entradas declarativas em YAML ao lado do código. Para workloads de agentes, o catálogo é a superfície de descoberta: um agente que precisa chamar um serviço consulta o catálogo, encontra a entrada, recupera a spec da API e a usa. O catálogo é a expressão operacional da camada de contexto da plataforma para informação de serviços.

Azure Red Hat OpenShift como cluster

O ARO é um serviço OpenShift gerenciado, operado conjuntamente por Microsoft e Red Hat. O control plane, o etcd e o API server são provisionados, atualizados e monitorados pela organização conjunta de engenharia sob um único SLA, tipicamente 99,95%, e uma única fatura da Microsoft cobre o cluster. Esse modelo operacional conjunto é uma das propriedades diferenciadoras do acelerador: incidentes que cruzam a fronteira Microsoft e Red Hat se resolvem dentro de um único relacionamento de suporte, em vez de entre dois fornecedores culpando um ao outro.

O ARO integra nativamente com a fundação Microsoft. O Entra ID é o provedor de identidade do cluster com RBAC governado por grupos Entra; Workload Identity emite credenciais federadas de curta duração, dispensando segredos de serviço de longa vida; o Azure Monitor ingere telemetria do cluster no plano unificado de observabilidade; o Azure Policy aplica governança na fronteira do cluster; e o Defender for Cloud fornece detecção de ameaças em runtime. Para workloads sensíveis a residência de dados, o ARO em Brazil South mantém cluster, workloads e serviços de fundação inteiramente na região, o que importa operacionalmente sob LGPD e BACEN.

> DEFINIÇÃO
>
> A escolha do cluster é uma escolha de modelo de suporte. ARO significa um SLA, uma fatura e um caminho de escalonamento cruzando Microsoft e Red Hat. Onde esse relacionamento único é requisito de compras, a decisão do cluster já está tomada.

OpenShift Pipelines e OpenShift GitOps

O OpenShift Pipelines é a distribuição Red Hat do Tekton, e sua propriedade estrutural é que pipelines são recursos Kubernetes, não estado de um orquestrador externo: um pipeline é um objeto YAML na API do cluster, observável pela mesma stack que observa os workloads. O OpenShift GitOps é a distribuição Red Hat do Argo CD, e faz do Git a única superfície de escrita do cluster: o estado é declarado no Git e reconciliado continuamente, então drift manual é detectado e revertido, e o cluster se autocorrige.

O padrão combinado é um laço: um desenvolvedor ou agente faz commit, o Pipelines constrói e valida, o Pipelines atualiza o repositório GitOps com a nova referência de imagem, o GitOps reconcilia o cluster, e a observabilidade fecha o laço. As políticas de sync são por ambiente: dev sincroniza automaticamente, staging exige aprovação explícita, produção exige sync explícito com múltiplos aprovadores. Isso é revisão manual aplicada na fronteira de deployment, e cada mudança é auditável no histórico do Git, o que satisfaz a postura de compliance com primitivas que o Git já tem.

> DEFINIÇÃO
>
> Duas propriedades fazem o trabalho pesado: produção não tem superfície de escrita direta, tudo flui pelo Git; e cada mudança é auditável no histórico. O padrão de incidente mais comum em produção, a mudança direta não autorizada, é prevenido estruturalmente.

Trusted Application Pipeline e a cadeia de suprimentos

O Red Hat Trusted Application Pipeline é a stack de segurança de cadeia de suprimentos integrada ao Pipelines. Ele gera SBOM completo a cada passo de build, assina cada artefato e atesta a proveniência do build em formato compatível com SLSA, escaneia dependências e artefatos produzidos em busca de vulnerabilidades, e aplica política que rejeita artefatos reprovados em build, deploy e runtime via admission control. O padrão em camadas com o GitHub é deliberado: GitHub Advanced Security no commit, TAP no build, Advanced Cluster Security no runtime.

Essa fiação responde diretamente a uma regressão medida: organizações que implantaram assistentes de codificação com IA antes de amadurecer sua postura de segurança de plataforma viram alta de 38% em vulnerabilidades exploráveis nos primeiros doze meses, segundo o Global Threat Report 2026 da CrowdStrike. Os quatro mecanismos do TAP fecham essa lacuna estruturalmente: dependências vulneráveis rejeitadas no build, artefatos não assinados rejeitados na admissão, divergências de proveniência rejeitadas no deploy, e SBOMs verificados continuamente contra o que de fato está rodando.

A camada de assistência de IA em três superfícies

O acelerador integra três superfícies de IA por audiência. O GitHub Copilot Enterprise assiste desenvolvedores na camada de código-fonte: completion, Copilot Chat e o coding agent do Copilot para trabalho multi-sessão. O Red Hat Lightspeed assiste operadores na camada de plataforma: operação de cluster, geração de playbooks Ansible, explicação de políticas. O Microsoft Foundry Agent Service roda os agentes customizados que o cliente constrói, com Entra Agent ID para identidade e observabilidade do Foundry para trajetórias. Desenvolvedores encontram o Copilot no IDE, operadores encontram o Lightspeed no console, e agentes de produção rodam no Foundry.

A escolha de design que vale notar é que a camada de runtime de agentes não varia por acelerador: agentes customizados no Three Horizons rodam na fundação Microsoft, o mesmo runtime que a variante Open Horizons usa, porque a Red Hat não entrega um runtime de agentes concorrente em 2026. A diferenciação entre os dois aceleradores vive nas camadas de IDP e cluster. A camada de primitivas de IA, gateway de modelos, runtime de agentes, avaliação e observabilidade, é fundação compartilhada.

Referências

Documentação de componente para cada elemento nomeado da stack.

- Documentação do Red Hat Developer Hub, o conjunto curado de plugins, software templates e ciclo de suporte.
- Documentação do Azure Red Hat OpenShift, o modelo operacional conjunto, SLA e integrações Azure.
- Documentação do OpenShift Pipelines, o substrato de CI/CD nativo de Kubernetes baseado em Tekton.
- Documentação do OpenShift GitOps, o modelo de reconciliação baseado em Argo CD e as políticas de sync.
- Red Hat Trusted Application Pipeline, SBOMs, assinatura, atestação e aplicação de política.
- SLSA, Supply-chain Levels for Software Artifacts, o framework de proveniência e atestação que o TAP implementa.

- Red Hat Lightspeed, a camada de assistência de IA do lado do operador no portfólio Red Hat.
- CrowdStrike, 2026 Global Threat Report, a regressão de segurança medida que a fiação de cadeia de suprimentos endereça.

Paula Silva

AI-NATIVE SOFTWARE ENGINEER

in/paulanunes

v1.0.0 · 2026-07-03

Building the future of software development with AI and Agentic DevOps

02 HORIZONTE 1

Otimizar o Presente: uma fundação Azure segura e observável nos primeiros 90 dias.

O Horizonte 1 é a fundação crível. Nos primeiros 90 dias o acelerador provisiona um ambiente Azure de nível de produção via 16 módulos Terraform: o cluster ARO, rede zero-trust, identidade e segredos, registry e bancos de dados, disaster recovery e governança de custos, e faz o onboarding dos três primeiros times por seis templates H1 de Golden Path. Tudo acima desta camada assume que ela foi bem feita, e é por isso que ela vem primeiro e nada nela é opcional.

O escopo dos primeiros 90 dias

O Horizonte 1 entrega a plataforma mínima viável: o cluster ARO provisionado com Workload Identity, Key Vault e integração com Azure Monitor; o Red Hat Developer Hub implantado, com marca aplicada e protegido por GitHub OAuth; o OpenShift GitOps com políticas de sync por ambiente; o OpenShift Pipelines; e o catálogo inicial de templates H1 carregado. A sequência de implantação roda o trabalho de fundação nas semanas zero a quatro e o onboarding de times da semana quatro em diante, com medições DORA ativas para cada time onboardado desde o início.

O marco que fecha o Horizonte 1 é observável, não cerimonial: o primeiro time de aplicação entrega um workload por um Golden Path, e seu lead time de PR a produção, medido na nova plataforma, supera o baseline anterior do próprio time. Isso normalmente acontece na semana seis a oito. Depois que o primeiro time entrega, os três seguintes fazem onboarding com menos atrito, porque os padrões de implantação estão validados e a documentação foi testada em campo.

Rede zero-trust

O módulo de rede provisiona uma rede virtual segmentada: subnets dedicadas para os nós master e worker do ARO, private endpoints e tráfego de gestão, com network security groups restringindo fluxos de entrada e saída em cada fronteira de subnet. Todos os serviços PaaS, Key Vault, o registry de contêineres e os bancos de dados, são alcançados exclusivamente por private endpoints resolvidos por zonas de Azure Private DNS. Nenhum serviço de dados tem exposição pública em qualquer configuração que o acelerador entrega.

A postura é deliberadamente preventiva em vez de detetiva: a superfície de ataque é reduzida antes de os workloads chegarem, não monitorada até tomar forma depois. Controle de egress, resolução privada e isolamento de subnets são geridos por Terraform, o que significa que as propriedades de segurança da rede são revisáveis em um pull request e reproduzíveis entre ambientes, em vez de viverem como estado de console que sofre drift.

Identidade, segredos e proteção de runtime

O módulo de segurança aplica uma regra em todo lugar: nenhuma credencial estática. Cada serviço usa managed identity; GitHub Actions e workloads Kubernetes autenticam por Workload Identity Federation sobre OIDC, então não existe segredo de longa vida para vazar; e as atribuições RBAC são de menor privilégio, com escopo por resource group. O Azure Key Vault, com soft delete e purge protection habilitados, é a fonte única de verdade para os segredos que precisam existir.

A proteção de runtime vem do Microsoft Defender for Containers, que observa o cluster ARO em busca de atividade suspeita de processos, escalação de privilégio e movimento lateral, escaneia imagens por vulnerabilidades e avalia o cluster contra benchmarks CIS. O RBAC completo integra Azure AD com OpenShift OAuth, e Security Context Constraints são aplicadas no nível do cluster. O modelo de identidade é a parte do Horizonte 1 em que a camada agêntica mais vai se apoiar depois: agentes recebem identidades do mesmo jeito que serviços.

> DEFINIÇÃO
>
> Uma regra, aplicada por construção: sem senhas, sem segredos de service principal em código, sem credenciais de longa vida em lugar nenhum. Managed identity mais Workload Identity Federation é o default, e a lista de exceções está vazia.

Registry e serviços de dados

O Azure Container Registry roda com geo-replicação para latência de pull multi-região, escaneamento de vulnerabilidades embutido no push, private endpoint em vez de exposição pública, e autenticação de pull por managed identity a partir do ARO, o que elimina credenciais de registry como classe de vazamento. O módulo de bancos provisiona PostgreSQL Flexible Server como backend do Developer Hub com backups automáticos e alta disponibilidade de produção, Redis para sessões e cache, e opcionalmente Cosmos DB atrás de uma flag Terraform para acesso distribuído globalmente.

Cada um desses serviços é alcançado apenas por private endpoints. A consequência prática para o time de plataforma é que as conversas de compliance da camada de dados ficam curtas: não há caminho público a enumerar, e a cadeia de verificação de SBOM a runtime do capítulo 01 se estende até as imagens que o registry admite.

Resiliência, custo e nomenclatura

O disaster recovery é explícito em vez de aspiracional. O Velero faz backup do estado do cluster e dos volumes persistentes para o Azure Blob Storage em agenda configurável; storage accounts e bancos replicam para uma região secundária com geo-redundância; e objetivos de ponto e tempo de recuperação são definidos por perfil de dimensionamento, do melhor esforço em dev a SLAs estritos de produção. Recuperação é uma propriedade testada da plataforma, não um documento.

A governança de custos chega no Horizonte 1 porque retrofitar depois nunca acontece. Orçamentos Azure carregam alertas de limiar roteados para e-mail e Teams, políticas de tagging são obrigatórias em todo recurso, e dashboards de análise de gasto aparecem no

Grafana ao lado das métricas de plataforma. Todos os recursos seguem as convenções de nomenclatura do Cloud Adoption Framework, mantendo o parque pesquisável e auditável entre ambientes e subscriptions.

> DEFINIÇÃO
>
> Seis templates H1 de Golden Path fazem o onboarding dos primeiros times: basic-cicd, documentation-site, infrastructure-provisioning, new-microservice, security-baseline e web-application. Cada um vem com template.yaml, código esqueleto e documentação.

Referências

Documentação de componente para a fundação do Horizonte 1.

- Documentação do Azure Red Hat OpenShift, arquitetura de cluster, clusters privados e autoscaling por MachineSet.
- Documentação do Azure Key Vault, soft delete, purge protection e acesso a segredos com escopo RBAC.
- Documentação do Microsoft Entra Workload ID, federação de identidade de workload sobre OIDC para pipelines e workloads.
- Microsoft Defender for Containers, detecção de ameaças em runtime e avaliação contra benchmarks CIS.
- Documentação do Velero, backup e restore de estado de cluster e volumes persistentes.
- Convenções de nomenclatura do Cloud Adoption Framework, o padrão de nomes aplicado pelo módulo CAF de naming.

Paula Silva

AI-NATIVE SOFTWARE ENGINEER

in/paulanunes

v1.0.0 · 2026-07-03

Building the future of software development with AI and Agentic DevOps

03 HORIZONTE 2

Capacidades Emergentes: a infraestrutura vira produto interno.

O Horizonte 2 é onde a infraestrutura crua vira algo que desenvolvedores de fato querem usar. Nos meses três a seis a plataforma ganha sua espinha GitOps em escala, sua porta de entrada no Developer Hub, seu catálogo completo de 22 templates de Golden Path, e a stack de observabilidade que torna a operação baseada em evidência. O critério de saída é adoção: pelo menos metade dos workloads elegíveis rodando pela plataforma, catálogo populado e TechDocs em uso na organização inteira.

GitOps em escala: app-of-apps e políticas de sync

O acelerador configura o Argo CD com o padrão app-of-apps: uma aplicação raiz gerencia todas as aplicações filhas, então a stack inteira da plataforma é implantada em ordem declarada por sync waves. As políticas de ambiente codificam o modelo de revisão em configuração, não em documentos de processo. Dev sincroniza automaticamente, mudanças entram na hora; staging é semiautomático, sincroniza com aprovação; produção é manual, exige aprovação humana explícita. O controle de acesso roda por SSO com Azure AD via Dex, com papéis de platform-admin e developer aplicados por RBAC.

Na escala do Horizonte 2, a disciplina que importa é estrutura de repositório. Parques multi-tenant produzem repositórios GitOps com centenas ou milhares de aplicações, e a relação aplicação-repositório vira a superfície operacional real. O acelerador entrega convenções opinadas de repositório exatamente por isso: o padrão sobrevive à escala apenas quando a estrutura é decidida antes de a quinquagésima aplicação chegar, não depois.

Developer Hub como porta de entrada

O Developer Hub é o ponto de entrada unificado de cada desenvolvedor: um lugar só para descobrir serviços, criar aplicações por scaffolding, navegar documentação e checar a saúde do sistema. O Software Catalog é o registro centralizado de cada serviço, API e componente; o TechDocs renderiza documentação versionada ao lado de cada entidade do catálogo; e a integração com GitHub fornece autenticação OAuth e acesso via GitHub App a repositórios e workflows.

O Horizonte 2 define metas explícitas de adoção para o portal porque portal sem adoção é mobília. Espera-se que o Service Catalog atinja pelo menos 80% de cobertura dos serviços reais, e que o TechDocs esteja em uso na organização inteira. Ambos são mensuráveis a partir do próprio catálogo, o que mantém a conversa honesta: ou o parque é descobrível pelo portal, ou não é.

> DEFINIÇÃO
>
> O teste do portal é cobertura, não existência. Um catálogo com 80% de cobertura muda como agentes e humanos descobrem serviços. Um catálogo com 20% ensina todo mundo a contornar o portal, e a plataforma perde a porta de entrada.

O catálogo de 22 templates de Golden Path

O Horizonte 2 libera o catálogo completo de Golden Paths: 22 templates pelos três horizontes, cada um com template.yaml, código esqueleto e documentação. Os nove templates H2 cobrem o meio do parque: api-gateway com rate limiting e políticas de autenticação, api-microservice com OpenAPI e validação, batch-job como CronJob com retries e alertas, data-pipeline ligado ao Event Hub, event-driven-microservice com integração a Service Bus ou Event Hubs, gitops-deployment para onboarding rápido, reusable-workflows como biblioteca compartilhada de Actions, ado-to-github-migration para migrações guiadas, e um template de microservice de uso geral.

A função do catálogo é padronização por conveniência, não por decreto. Quando o caminho sancionado também é o mais rápido, os times o escolhem voluntariamente, e cada serviço criado por template chega com CI/CD, linha de base de segurança, observabilidade e registro no catálogo já ligados. A meta de saída do Horizonte 2 é

adoção da plataforma em 50% ou mais dos workloads elegíveis, medida, como tudo aqui, pelos dados da própria plataforma.

A stack de observabilidade

A stack é Prometheus, Grafana, Alertmanager e Jaeger, implantados e configurados pelo acelerador. O Prometheus roda duas réplicas com retenção de 15 dias e 50GB de storage; o Grafana autentica por SSO do Azure AD e entrega três dashboards curados: Platform Overview como centro de comando diário do time de plataforma, Cost Management com atribuição por time e namespace, e Golden Path Application com os quatro sinais dourados de cada serviço criado por template. O Alertmanager roteia para PagerDuty e Microsoft Teams com regras de silêncio e inibição, e o Jaeger rastreia requisições entre serviços.

Mais de 30 regras de alerta vêm por padrão, cobrindo cinco camadas: alertas de infraestrutura para pressão de nós, disco, rede e saúde de componentes do ARO; alertas de aplicação para SLOs de taxa de erro, limiares de latência, crash loops e deployments pendentes; alertas de workload de IA para cota de tokens, latência de inferência, disponibilidade de endpoints de modelo e falhas de avaliação; alertas de GitOps para falhas de sync e aplicações fora de sincronia; e alertas de segurança para achados do Defender, violações de política, acessos incomuns e expiração de certificados. O grupo de alertas de IA existe no Horizonte 2 de propósito: a plataforma é instrumentada para agentes antes de os agentes chegarem.

Segredos, runners e ingress

Três serviços operacionais completam o horizonte. O External Secrets Operator sincroniza continuamente segredos do Azure Key Vault para Kubernetes Secrets via ClusterSecretStore, com rotação automática e zero segredos guardados em Git ou variáveis de ambiente. Runners self-hosted do GitHub no ARO dão aos workflows de CI/CD acesso a recursos privados, ACR, Key Vault, o próprio cluster, sem exposição pública; os runners escalam pela profundidade da fila, autenticam por managed identity e são efêmeros, então cada job parte de estado limpo.

Ingress e TLS ficam com o NGINX Ingress Controller com roteamento por caminho e host, rate limiting e capacidade de WAF; o cert-manager provisiona e renova certificados do Let's Encrypt sem gestão manual; e o External DNS cria e atualiza registros no Azure DNS conforme recursos de ingress aparecem e mudam. Nenhum desses três serviços é glamouroso, e os três são a diferença entre uma plataforma que se opera sozinha e uma que aciona o pager do time.

> DEFINIÇÃO
>
> O fluxo de segredos em uma linha: o Azure Key Vault é a fonte de verdade, o External Secrets Operator sincroniza e rotaciona, os pods consomem Kubernetes Secrets em runtime. Nada sensível vive em Git, em variável de ambiente ou em definição de pipeline.

Referências

Documentação de componente para os serviços de plataforma do Horizonte 2.

- Documentação do Argo CD, o padrão app-of-apps, sync waves e políticas de sync.
- Documentação do Backstage, as fundações de catálogo, TechDocs e scaffolder sob o Developer Hub.
- Documentação do Prometheus, coleta de métricas e o modelo de regras de alerta.
- Documentação do Grafana, dashboards e single sign-on com Azure AD.
- Documentação do External Secrets Operator, o padrão ClusterSecretStore e a rotação de segredos.
- Documentação do cert-manager, provisão e renovação automatizadas de certificados TLS.
- Documentação do Jaeger, tracing distribuído entre os serviços da plataforma.

Paula Silva

AI-NATIVE SOFTWARE ENGINEER

in/paulanunes

v1.0.0 · 2026-07-03

Building the future of software development with AI and Agentic DevOps

04 HORIZONTE 3

Transformar o Futuro: a camada agêntica pousa numa plataforma pronta.

O Horizonte 3 é a razão de os outros dois existirem. Nos meses seis a doze a plataforma ganha sua infraestrutura de IA como código, seus sete Golden Paths de IA, sua frota de 17 agentes especializados de Copilot Chat, e a maquinaria de avaliação e observabilidade que torna a autonomia governável. O critério de saída é um processo de negócio real rodando de ponta a ponta por um agente, com revisão humana nas fronteiras de promoção e cada trajetória, token e custo contabilizados.

Azure AI Foundry, provisionado como código

O módulo de AI Foundry provisiona a infraestrutura completa de IA como Terraform: o hub do Foundry, projetos, deployments de modelos e o Azure AI Search, reproduzíveis e versionados como qualquer outra camada da plataforma. Os deployments de referência incluem GPT-4o para trabalho multimodal, o3 para raciocínio mais difícil e previews de próxima geração conforme chegam à subscription, todos expostos por managed online endpoints para inferência e pelo gateway do Foundry para roteamento, medição e política.

A propriedade estrutural que vale internalizar é que a infraestrutura de IA deixa de ser especial. Deployments de modelo vivem no mesmo fluxo de pull request que regras de rede; o vector store é provisionado pelo mesmo pipeline que o banco; e rollbacks funcionam do mesmo jeito em todo lugar. Times pedem capacidade de IA por Golden Paths, não por tickets, e a plataforma mede o que consomem por workload desde o primeiro dia.

Os sete Golden Paths de IA

O Horizonte 3 entrega sete templates focados em IA. O rag-application cria por scaffolding uma aplicação completa de geração aumentada por recuperação: ingestão de documentos, embeddings vetoriais no Azure AI Search e interface de chat, com

observabilidade embutida. O multi-agent-system fornece um padrão de orquestrador com subagentes especialistas, integrações de ferramentas MCP, protocolos de handoff e estado compartilhado. O mlops-pipeline cobre o ciclo de ML completo de treino a avaliação, registro, deploy e monitoramento de drift. O foundry-agent cria aplicações de agente único no Foundry Agent Service com identidade e avaliação ligadas.

Os três restantes fecham o laço de qualidade e operação: o ai-evaluation-pipeline transforma avaliação de modelos em pipeline reutilizável, o copilot-extension cria extensões customizadas do GitHub Copilot apoiadas em serviços de IA do Azure, e o sre-agent-integration liga triagem de incidentes assistida por IA, execução de runbooks e análise de causa raiz ao fluxo operacional da plataforma. Juntos, os sete templates significam que um time de aplicação pode subir um workload de IA governado na mesma tarde em que decide construí-lo, e cada um desses workloads chega observável e avaliável por construção.

> DEFINIÇÃO
>
> Sete templates, uma propriedade: um workload de IA criado por Golden Path chega com identidade, observabilidade, avaliação e medição de custo já acopladas. O time escreve a lógica de domínio. A plataforma fornece tudo que o torna nível de produção.

A frota de 17 agentes e o MCP

O acelerador entrega 17 agentes especializados de Copilot Chat que operam direto no VS Code, cobrindo fluxos de desenvolvimento, deploy, operação e gestão de incidentes, apoiados em 16 skills operacionais e 13 integrações de servidores MCP que expõem capacidades da plataforma como ferramentas. Cada agente tem papel definido, acesso explícito a ferramentas e fronteiras comportamentais em três níveis: ações que SEMPRE executa, ações em que precisa PERGUNTAR ANTES, e ações que NUNCA executa. O modelo de fronteiras é a primitiva de governança: autonomia é concedida por classe de ação, não por agente.

A frota também é como o time de plataforma come da própria comida. Os primeiros agentes adotados numa implantação Three Horizons são tipicamente internos à

plataforma: triagem de incidentes, autoria de runbooks, revisão de políticas, remediação de drift. Essa adoção interna, pelo menos uma tarefa recorrente da plataforma absorvida por um agente, é metade do teste de maturidade que este corpo de trabalho chama de Mandato Duplo, e a expectativa é que passe por volta da semana seis da implantação.

Gates de avaliação e observabilidade de agentes

Avaliação é gate de deploy, não relatório. O pipeline de avaliação pontua saídas de modelo por relevância, coerência e fundamentação contra datasets de referência, filtra conteúdo nocivo, enviesado ou violador de política, e alimenta os resultados no CI/CD como gates: uma avaliação reprovada bloqueia o deployment automaticamente, do mesmo jeito que um teste reprovado bloqueia. A promoção de um agente para um raio de ação maior segue as mesmas fronteiras de revisão manual de qualquer mudança de produção.

A observabilidade se estende a tudo que os agentes fazem: traces de trajetórias, custo por requisição e por workload, amostragem de saídas para revisão de qualidade e resultados de avaliação ao longo do tempo, ao lado dos alertas de workload de IA do Horizonte 2. O dashboard de AI Usage vira superfície primária de KPI da plataforma. A saída do Horizonte 3 é concreta: pelo menos um processo de negócio real roda de ponta a ponta por um agente, com revisão humana nas fronteiras de promoção, em 90 a 180 dias do início da implantação.

> DEFINIÇÃO
>
> A postura de governança em uma frase: agentes são workloads de primeira classe com identidade, orçamento, gates de avaliação e auditoria, e a autonomia se expande apenas conforme a evidência acumula. É isso que torna o futuro de 100 agentes por humano operável.

Referências

Documentação de componente para a camada agêntica do Horizonte 3.

- Documentação do Azure AI Foundry, o hub, projetos, deployments de modelo e gateway.
- Documentação do Azure AI Search, a camada de recuperação vetorial sob o Golden Path de RAG.
- Microsoft Agent Framework 1.0, o framework sobre o qual os agentes customizados são construídos.
- Model Context Protocol, o padrão de integração de ferramentas por trás dos 13 servidores MCP.
- Documentação do GitHub Copilot, Copilot Chat, agentes customizados e extensões.
- Convenções semânticas GenAI do OpenTelemetry, o padrão de telemetria para trajetórias e custo de agentes.

Paula Silva

AI-NATIVE SOFTWARE ENGINEER

in/paulanunes

v1.0.0 · 2026-07-03

Building the future of software development with AI and Agentic DevOps

05 A DECISÃO

Escolha pelas razões certas, implante na ordem certa.

O Three Horizons não é a resposta certa para todo cliente, e fingir o contrário desperdiça o ano de todo mundo. Este capítulo de fechamento dá estrutura à decisão: o que o acelerador materializa em termos de governança, o que custa comercialmente e o que o suporte conjunto compra, quais perfis de cliente pendem a favor e contra, e a sequência de três fases que leva uma decisão assinada a um marco de primeiro valor medido.

O que o crosswalk de pilares compra

Em termos de governança, o acelerador materializa os quatro pilares CNCF com componentes nomeados. Golden Paths são os Software Templates do RHDH com CI/CD Tekton gerado, deployment via Argo CD e infraestrutura Terraform. Guardrails são as políticas do Advanced Cluster Security, admission controllers, Azure Policy nas fronteiras de cloud e cluster, Workload Identity no lugar de credenciais de longa vida, e assinatura TAP na admissão. Safety Nets são a reconciliação GitOps, o plano unificado de observabilidade, backup Velero, entrega progressiva via service mesh e detecção de runtime do ACS. Revisão Manual é o sync com gates por ambiente mais a promoção governada de agentes.

A propriedade diferenciadora é que esses pilares chegam pré-integrados no deploy. O cliente não parte de um cluster em branco para integrar cada pilar; parte de uma fundação integrada e customiza política e conteúdo. Esse trabalho de integração absorvido é a explicação estrutural da compressão de 9 a 18 meses para 90 a 180 dias: a parte mais lenta da construção de plataforma é exatamente a parte que o acelerador entrega pronta.

A estrutura comercial e o suporte conjunto

Três linhas comerciais compõem o acelerador. Subscriptions Red Hat cobrem RHDH, ARO, Pipelines, GitOps, Service Mesh, ACM, ACS, TAP e Lightspeed onde licenciado.

Consumo Azure cobre computação e storage do cluster mais os serviços de fundação: Key Vault, AI Foundry, Monitor, Entra. GitHub Enterprise cobre licenciamento por seat com Advanced Security e Copilot Enterprise. Comparando com a variante open-source desta arquitetura, a linha de subscriptions Red Hat é o principal diferenciador de custo, e o contrato de suporte é exatamente o que essa linha compra.

O modelo de suporte conjunto é a razão comercial mais forte pela qual clientes escolhem este caminho. Microsoft e Red Hat operam o ARO conjuntamente, com caminhos de escalonamento entre si, então um incidente que cruza a fronteira é acionado por uma parte e resolvido internamente, em vez de estagnar enquanto dois fornecedores se culpam. Para algumas organizações de compras esse relacionamento único é requisito inegociável, e onde é, a decisão está efetivamente tomada antes de qualquer comparação técnica começar.

> DEFINIÇÃO
>
> Leia a estrutura de custo como troca, não como penalidade: a linha de subscriptions Red Hat compra o contrato de suporte conjunto, a postura de certificação e o ciclo de vida enterprise. Clientes que não precisam dessas três coisas não deveriam pagar por elas, e a lógica do capítulo diz isso explicitamente.

Perfis de cliente: a favor, contra e indecisos

O perfil que pende para o Three Horizons tem pelo menos três destas propriedades: investimento Red Hat prévio em escala material, RHEL, OpenShift ou Ansible; requisitos regulatórios ou de auditoria em que certificações FedRAMP, FIPS ou Common Criteria bloqueiam compras; uma estratégia de fornecedores que prefere um relacionamento conjunto forte a vários frouxos; e concentração em serviços financeiros, energia, governo ou saúde. As implantações de referência por trás deste playbook estão exatamente nesses setores, e a biblioteca de políticas vem opinada para LGPD, BACEN, ANEEL e ANS por causa disso.

O perfil que pende contra é igualmente claro: nenhuma pegada Red Hat prévia, estratégia de minimizar contagem de fornecedores, forte expertise interna em Kubernetes upstream, ou cultura de engenharia OSS-first que valoriza transparência e ritmo acima de suporte de

fornecedor. Esses clientes são melhor servidos pela variante open-source Open Horizons em AKS e Backstage upstream. Quando a liderança está dividida ou as duas culturas coexistem, rode uma descoberta comparativa de um dia em vez de discutir: pontue o parque contra essas propriedades de perfil e deixe a evidência recomendar.

A sequência de implantação em três fases

Fase um, semanas zero a quatro, é fundação: o cluster ARO provisionado com Workload Identity, Key Vault e integração com Monitor; RHDH implantado, com marca e protegido por GitHub OAuth; GitOps com sync por ambiente; Pipelines; e o catálogo de templates H1 carregado. Fase dois, semanas quatro a doze, é onboarding: os três primeiros times de aplicação chegam pelos templates H1, o Service Catalog popula, o TechDocs é adotado, políticas ACS se aplicam, o TAP se integra ao Pipelines, e a medição DORA fica ativa para cada time onboardado.

Fase três, meses três a doze, é escala: templates H2 lançam e são adotados, o dashboard de custos entra no ar, o primeiro agente de Copilot Chat é implantado, a cobertura do catálogo cruza 80%, multi-cluster expande via ACM onde aplicável, e os templates H3 chegam aos early adopters. O marco que ancora a sequência inteira chega na semana seis a oito: o primeiro time entrega por um Golden Path com lead time mais rápido que seu próprio baseline anterior. Esse número, não uma demo, é o que deve embasar decisões de expansão.

> DEFINIÇÃO
>
> O marco de primeiro valor é deliberadamente pouco impressionante: um time, um workload, um número de lead time melhor que o próprio baseline. Tudo que escala depois é construído sobre esse fato verificado, e é por isso que a sequência se recusa a pulá-lo.

Mapeamento de maturidade e o teste permanente

A escada H1/H2/H3 é um modelo de sequenciamento; o CNCF Platform Maturity Model é um modelo de medição, e os dois se compõem. O Horizonte 1 corresponde

aproximadamente ao nível Operational da CNCF na maioria das dimensões, o Horizonte 2 ao Scalable, e o Horizonte 3 ao Optimizing ou AI-Native. Usados juntos, a escada diz ao time de plataforma o que construir a seguir e o modelo CNCF diz quão maduro o resultado de fato é, o que mantém as conversas de roadmap ancoradas em estado avaliado em vez de entusiasmo.

O teste permanente de uma implantação saudável é o Mandato Duplo, e no Three Horizons ele é tratável por volta da semana seis: pelo menos uma tarefa recorrente do próprio time de plataforma foi absorvida por um agente, e pelo menos um time de aplicação consome o gateway de modelos por um Golden Path. Quando as duas metades valem, a plataforma come da própria comida e serve seus clientes ao mesmo tempo, e a camada agêntica que vem depois pousa sobre capacidade demonstrada em vez de um slide.

Referências

Fontes primárias para o framework de decisão, o modelo comercial e o mapeamento de maturidade.

- Documentação do Azure Red Hat OpenShift, a operação conjunta, o SLA e o modelo de suporte em que o argumento comercial se apoia.
- CNCF Platform Engineering Maturity Model, o modelo de medição de cinco dimensões sobre o qual os horizontes se mapeiam.
- CNCF Platforms Whitepaper, as definições de pilares por trás do crosswalk.
- Red Hat Advanced Cluster Security, o componente de segurança de runtime das linhas de Guardrails e Safety Nets.
- Red Hat Advanced Cluster Manager, política e governança multi-cluster para a fase de escala.
- Programa de pesquisa DORA 2025, as métricas de entrega que ancoram o marco de primeiro valor e a decisão de expansão.

Paula Silva

AI-NATIVE SOFTWARE ENGINEER

in/paulanunes

v1.0.0 · 2026-07-03

Building the future of software development with AI and Agentic DevOps