Controles de segurança de IA generativa

Information Security for Workload Factory

NetApp
September 16, 2026

Índice

Controles de segurança de IA generativa

Saiba mais sobre controles de IA no NetApp Workload Factory

O NetApp Workload Factory inclui recursos de IA generativa que aceleram o diagnóstico, ao mesmo tempo que reforçam os controles de segurança para acesso ao modelo, escopo de dados e comportamento em tempo de execução.

O Workload Factory ativa o diagnóstico por IA por padrão. Você pode desativar os recursos de IA generativa a qualquer momento nas configurações de **Administração** do Workload Factory. "Saiba mais sobre como gerenciar os recursos do GenAI."

Visão geral dos recursos de IA com foco em segurança

Atualmente, o Workload Factory aplica IA generativa aos seguintes recursos:

- Assistente "Pergunte-me" (RAG com documentação selecionada da NetApp e respostas baseadas em fontes)
- Análise de latência
- Análise de eventos EMS
- Análise do log de erros

Visão geral do AWS Bedrock para diagnósticos de IA

O Workload Factory usa o Amazon Bedrock para inferência de IA. A inferência é executada na sua conta AWS, com suas credenciais e perfil de inferência configurados.

Informações relacionadas

- "Opt-in do Amazon Bedrock para diagnósticos de IA do NetApp Workload Factory"
- "Limitações das ferramentas de IA para o NetApp Workload Factory"
- "Parâmetros do modelo de IA e faturamento para o NetApp Workload Factory"

Opt-in do Amazon Bedrock para diagnósticos de IA do NetApp Workload Factory

NetApp Workload Factory AI diagnostics utiliza o Amazon Bedrock para analisar eventos e o recurso está habilitado por padrão. Antes que o Workload Factory possa invocar o modelo Bedrock selecionado e coletar dados de diagnóstico, você deve configurar as permissões do IAM, um perfil de inferência e um link Lambda conectado com capacidade SSH. Se algum desses componentes estiver ausente ou a configuração do Bedrock for removida, os diagnósticos de IA ficarão indisponíveis.

O diagnóstico por IA está ativado por padrão.

Antes de começar

- Certifique-se de que sua conta inclua um perfil de inferência do AWS Bedrock.
- Tenha um link Lambda conectado com capacidade SSH para coleta de dados diagnósticos.

Passos

1. Conceda permissões do Bedrock à sua função do IAM:
    - `bedrock:InvokeModel`
    - `bedrock:InvokeModelWithResponseStream`
    - `bedrock:ListInferenceProfiles`
    - `bedrock:GetInferenceProfile`
2. Configure o ARN do perfil de inferência nas configurações do o Workload Factory.

Caso falte algum pré-requisito, o sistema informa o componente específico que está ausente e os recursos de IA permanecem inativos.

Desativar diagnósticos de IA

Para desativar o diagnóstico por IA:

- Remova as permissões do Bedrock da sua função IAM, ou
- Exclua a configuração do Bedrock nas configurações do Workload Factory.

Os diagnósticos por IA tornam-se imediatamente indisponíveis e nenhum processamento de dados residual ocorre.

Perfis de inferência regional e entre regiões

A inferência do Bedrock é executada na região da AWS especificada pelo seu perfil de inferência. Se você usar um perfil de inferência entre regiões, a AWS pode rotear a solicitação para qualquer região na configuração do perfil, conforme o comportamento padrão do AWS Bedrock. Os dados não saem da infraestrutura da AWS.

Limitações das ferramentas de IA para o NetApp Workload Factory

NetApp Workload Factory impõe limites rigorosos sobre como os recursos de IA interagem com seus ambientes AWS e ONTAP, para que você possa confiar nos diagnósticos assistidos por IA sem correr o risco de alterações indesejadas. O Amazon Bedrock usa ferramentas de diagnóstico somente leitura da AWS CLI e da ONTAP CLI que bloqueiam comandos de modificação e limitam a saída por tamanho e execução, e todos os dados utilizados são apenas transitórios e não são retidos pelo NetApp Workload Factory ou pela AWS.

Controles de segurança da ferramenta

Ferramenta	Capacidade	Controles de segurança
AWS CLI (somente leitura)	Execute comandos somente leitura da AWS CLI (describe, list, get)	Bloqueia todos os verbos de mutação (criar, excluir, modificar, atualizar, inserir, remover). Máximo 10 comandos por etapa. Saída truncada em 20 KB.
CLI do ONTAP (somente leitura)	Execute comandos ONTAP CLI somente leitura usando SSH	Bloqueia todos os verbos de mutação (excluir, destruir, criar, modificar, mover). Saída truncada.

O agente não pode modificar, criar ou excluir recursos. Ele só pode observar e diagnosticar.

Limites de retenção de dados e treinamento de modelos

De acordo com a política da AWS, o Amazon Bedrock não armazena nem utiliza suas entradas e saídas para treinar ou aprimorar modelos de base. Para inferência sob demanda (usada pelo o Workload Factory), a AWS processa seus dados de forma transitória e não os retém após retornar a resposta.

Parâmetros do modelo de IA e faturamento para o NetApp Workload Factory

NetApp Workload Factory define a configuração do modelo e os limites de uso e faturamento relevantes para a governança de IA. O modelo configurado usa parâmetros determinísticos e retorna JSON estruturado para oferecer resultados de análise consistentes. A AWS cobra a inferência do Amazon Bedrock na sua conta, enquanto o Workload Factory monitora o uso de IA e fornece visibilidade do custo mensal por conta. Estas seções descrevem os parâmetros do modelo e os limites que se aplicam ao consumo de IA.

Modelo e parâmetros

Parâmetro	Valor
Modelo	Anthropic Claude (usando um perfil de inferência entre regiões)
Temperatura	0 (determinístico, sem aleatoriedade)
Tokens máximos de saída	2.048
Etapas máximas de raciocínio	15 por análise
Formato de saída	JSON estruturado (resumo, gravidade, causa raiz, ações corretivas)

Limites de faturamento e uso

- A AWS cobra a inferência do Bedrock na sua conta (seu perfil de inferência, suas credenciais).
- O Workload Factory monitora internamente o uso de IA.

- O Workload Factory oferece visibilidade mensal dos custos por conta.

  "Acompanhe os custos de armazenamento no NetApp Workload Factory"

Assistente Ask Me AI

Pergunte-me sobre a arquitetura de segurança do NetApp Workload Factory

O Ask Me é um assistente generativo com inteligência artificial integrado ao NetApp Workload Factory. Ele oferece suporte conversacional em tempo real para Amazon FSx for NetApp ONTAP e tópicos do Workload Factory. As respostas são baseadas em documentação verificada da NetApp, e o Ask Me não pode acessar a internet pública nem usar dados do cliente para treinar o modelo. Múltiplas camadas de segurança ajudam a bloquear consultas maliciosas, limitar as respostas a domínios suportados e impedir que a entrada do usuário substitua as instruções do sistema. Quando o Ask Me usa ferramentas, limites de autorização, imposição de consultas somente leitura, um ambiente de teste temporário e registro de auditoria ajudam a restringir e monitorar a execução.

Você pode desativar o recurso "Pergunte-me" a qualquer momento nas configurações de **Administração** do NetApp Workload Factory, o que desativa o assistente para sua conta de usuário. "Saiba mais sobre como gerenciar os recursos do GenAI."

Arquitetura RAG em Ask Me

O Ask Me utiliza geração aumentada por recuperação (RAG).

- Base de conhecimento selecionada: as respostas são fundamentadas em um conjunto gerenciado de documentação verificada e publicada da NetApp.
- Pontuação e filtragem de recuperação: apenas o conteúdo suficientemente relevante é incluído no contexto do modelo.
- Sem acesso à internet: o modelo não consegue buscar, navegar ou extrair conteúdo externo.
- Atribuição da fonte: as respostas incluem referências aos documentos de origem utilizados.

Segurança de prompt em várias camadas

- Camada 1: O classificador de segurança (LLM como juiz) bloqueia consultas maliciosas (injeção de prompts, escalonamento de privilégios, exfiltração de dados, engenharia social, violações de políticas).

  O sistema registra as consultas bloqueadas, e o modelo generativo nunca as vê.

- Camada 2: O reforço da segurança dos comandos do sistema impede que a entrada do usuário substitua as instruções do sistema.
- Camada 3: o escopo do domínio limita as respostas aos tópicos do Amazon FSx for NetApp ONTAP e do NetApp Workload Factory.
- Camada 4: a governança de uso de ferramentas valida parâmetros e executa consultas em um ambiente de teste reforçado; o modelo não pode executar operações arbitrárias.

Isolamento e privacidade dos dados do modelo

- Nenhum treinamento de modelo em dados de clientes
- Isolamento no nível da solicitação (sem vazamento de dados entre locatários)
- Contexto de conversação com escopo de sessão e histórico delimitado
- Sem memória de modelo persistente
- Infraestrutura de inferência gerenciada

Uso de ferramentas e segurança agente

- Ferramentas com escopo de autorização (limites de permissão do usuário aplicados)
- Tempos limite rigorosos para execução de ferramentas
- Sandbox efêmero com escopo de sessão para dados recuperados
- Imposição de consultas somente leitura com allowlisting
- Auditabilidade de chamadas de ferramentas com limpeza de parâmetros sensíveis

Controle de acesso para o NetApp Workload Factory

No NetApp Workload Factory, o Ask Me prioriza a segurança de acesso em sessões autenticadas, a responsabilização em nível de usuário e o tratamento de segredos protegidos durante as operações de tempo de execução. Os controles de autenticação validam cada sessão e associam as interações a um usuário específico. As credenciais confidenciais são recuperadas de um cofre de segredos gerenciado em tempo de execução e ocultadas dos registros. O serviço também aplica controles de infraestrutura, incluindo a execução de contêineres sem privilégios de root e uma lista de permissões CORS restrita.

Autenticação

- Autenticação JWT assinada com validação criptográfica (incluindo verificações de público, emissor e algoritmo, como RS256)
- Autenticação aplicada por middleware no limite do serviço
- Definição do escopo da identidade do usuário para que as interações sejam atribuídas a um usuário específico

Proteção de credenciais e manuseio de segredos

- Armazenamento seguro em cofre: as credenciais confidenciais usadas pelo serviço são armazenadas em um cofre de segredos gerenciado e recuperadas em tempo de execução. Nenhum segredo é armazenado no código do aplicativo, nos arquivos de configuração ou nas imagens do contêiner.
- Limpeza de logs: valores sensíveis (credenciais, tokens, senhas, chaves de acesso, certificados) são removidos dos logs antes de serem gravados.

Segurança da infraestrutura

- Execução de contêiner sem privilégios de root
- CORS restrito a uma lista explícita de domínios confiáveis da NetApp

Modos de execução Ask Me para NetApp Workload Factory

NetApp Workload Factory oferece múltiplos modos de execução para o Ask Me, cada um com diferentes características de segurança e privacidade. Quando o Ask Me utiliza integrações de ferramentas, por exemplo, consultando os recursos do próprio Workload Factory do cliente, a execução da ferramenta é controlada por meio de camadas de proteção. A execução habilitada para ferramentas opera dentro das permissões do usuário autenticado e aplica limites à duração e ao escopo de cada operação. Os dados recuperados permanecem em um ambiente isolado de sessão efêmero que bloqueia o acesso externo e é destruído ao final da sessão. Controles independentes de prompt e código impõem consultas somente leitura, enquanto registros de auditoria capturam a atividade da ferramenta com valores sensíveis removidos.

Salvaguardas de execução da ferramenta

- As ferramentas com escopo de autorização operam dentro das permissões do usuário autenticado.
- O tempo limite de execução da ferramenta limita operações de longa duração.
- O ambiente de dados efêmeros armazena dados recuperados por ferramentas no escopo da sessão, bloqueia acesso externo, E/S de arquivos, chamadas de rede e carregamento de extensões no nível do mecanismo e é destruído quando a sessão termina.
- A aplicação de consultas somente leitura valida as consultas geradas por meio de uma lista de permissões rigorosa.
- A segurança imposta por prompt e a segurança imposta por código são aplicadas de forma independente.
- Os registros de auditoria de chamadas de ferramentas incluem o nome da ferramenta, parâmetros, carimbos de data/hora e status do resultado, com remoção de valores confidenciais.

Controles de saída

- Aplicação do limite de tokens
- Saída determinística para operações de classificação/segurança crítica (temperatura 0)
- Transmissão com encerramento antecipado e limpeza

Pergunte-me sobre privacidade e disponibilidade para o NetApp Workload Factory

No NetApp Workload Factory, os controles de privacidade do Ask Me limitam a exposição dos dados, mantêm o processamento da sessão isolado e preservam os limites de privacidade para as interações do usuário. Suas entradas são processadas por um serviço de IA gerenciado que não as utiliza para treinar ou aprimorar modelos básicos. Os dados operacionais recuperados permanecem em armazenamento efêmero na memória durante a sessão e não são persistidos nem compartilhados. Uma cadeia de fallback multimodelo em diversas regiões da nuvem garante a disponibilidade quando o modelo principal está com desempenho limitado ou indisponível, enquanto os mesmos controles de segurança se aplicam aos modelos de fallback.

Tratamento e privacidade de dados

- Dados do usuário em prompts: processados por um serviço de IA gerenciado que não utiliza suas entradas para treinar ou aprimorar foundation models.
- Conteúdo da base de conhecimento: somente documentação NetApp selecionada e publicada; seus dados não estão incluídos.
- Dados operacionais: recuperados somente quando um usuário consulta seus próprios recursos; mantidos em armazenamento efêmero na memória durante a sessão e não persistidos ou compartilhados.
- Registro de auditoria: metadados de consulta (ID de usuário anonimizado, registros de data e hora, durações) registrados, com os campos sensíveis removidos.
- Dados de feedback: armazenados separadamente e vinculados a um ID de consulta, não a informações de identificação pessoal além do ID de usuário anonimizado.

Controles de disponibilidade e isolamento

O Ask Me emprega uma cadeia de fallback multimodelo em várias regiões de nuvem.

Se o modelo principal estiver com desempenho limitado ou indisponível, o sistema pode recorrer a modelos alternativos.

Todos os modelos estão sujeitos aos mesmos controles de segurança, reforço da segurança do sistema e garantias de isolamento.

Informações sobre direitos autorais

Informações sobre marcas comerciais