Planejamento para o sucesso MLOps

AWS Orientação prescritiva

AWS Orientação prescritiva: Planejamento para o sucesso MLOps

Table of Contents

Planejamento para o sucesso MLOps

Bruno Klein, Amazon Web Services ()AWS

Dezembro de 2021 (histórico do documento)

A implantação de soluções de aprendizado de máquina (ML) na produção apresenta muitos desafios que não surgem em projetos de desenvolvimento de software padrão. As soluções de ML são mais complexas e difíceis de acertar em primeiro lugar. Eles também existem em ambientes geralmente voláteis, onde a distribuição de dados se desvia significativamente ao longo do tempo por vários motivos esperados e inesperados.

Esses problemas são ainda mais agravados pelo fato de muitos profissionais de ML não terem formação em engenharia de software, portanto, talvez não estejam familiarizados com as melhores práticas desse setor, como escrever código testável, modularizar componentes e usar o controle de versão de forma eficaz. Esses desafios criam dívidas técnicas e as soluções se tornam mais complexas e difíceis de manter com o tempo, impulsionadas por um efeito combinado, para as equipes de ML.

Este guia enumera as melhores práticas de operações de ML (MLOps) que ajudam a mitigar esses desafios em projetos e cargas de trabalho de ML.

Por ser MLOps uma preocupação transversal, esses problemas afetam não apenas os processos de implantação e monitoramento, mas todo o ciclo de vida do modelo. Neste guia, as MLOps melhores práticas são organizadas em quatro áreas principais:

- Dados
- Treinamento
- Implantação
- Monitoramento

Resultados de negócios desejados

Implantar modelos de ML na produção é uma tarefa que exige esforço contínuo e uma equipe dedicada para manter esses recursos durante toda a vida útil (em alguns casos, até anos). Os modelos de ML podem gerar um valor considerável dos dados corporativos, mas têm altos custos. Para minimizar os custos, as empresas devem seguir as boas práticas em desenvolvimento de

software e ciência de dados. Eles devem estar cientes das nuances dos sistemas de ML, como o desvio de dados, que faz com que os modelos funcionem inesperadamente depois de um tempo. Ao estarem cientes dessas preocupações, as empresas podem atingir suas metas de negócios com segurança e agilidade no curto e longo prazo.

Existem vários tipos de modelos de ML, e os setores que eles visam têm vários tipos de tarefas de ML e problemas de negócios, então você precisa considerar um conjunto diferente de preocupações para cada modelo e setor. As práticas descritas neste guia não são específicas de um modelo ou negócio, mas se aplicam a um amplo conjunto de modelos e setores para melhorar os tempos de implantação, gerar maior produtividade e criar governança e segurança mais fortes.

Colocar modelos em produção é uma tarefa multidisciplinar que exige cientistas de dados, engenheiros de aprendizado de máquina, engenheiros de dados e engenheiros de software. Ao criar sua equipe de ML, recomendamos que você se concentre nessas habilidades e experiências.

Dados

DevOps é uma prática de engenharia de software que lida com a operacionalização de software. Os elementos comuns DevOps são código controlado por versão, pipelines de integração contínua e entrega contínua (CI/CD), testes unitários e criação e implantação de código reproduzíveis, que envolvem código. Os modelos de ML são um produto de código e dados, portanto, os dados precisam atender aos mesmos padrões do código. MLOps deve abordar questões relacionadas a dados, como manter a qualidade dos dados, como identificar casos extremos nos dados, como proteger os dados e como tornar os dados mais fáceis de manter.

Tópicos

Rótulo

Forneça instruções claras de rotulagem

Um conjunto de dados pode incluir amostras ambíguas que resultam em rotulagem inconsistente em todo o conjunto de dados. Por exemplo, considere a tarefa de rotular imagens que contenham um cachorro. Algumas amostras podem conter apenas um vislumbre do animal. Eles devem ser marcados com um rótulo positivo ou negativo? Esse tipo de problema pode ser resolvido fornecendo instruções claras e objetivas aos rotuladores.

Use a votação majoritária

Agora, considere a questão de rotular um speech-to-text conjunto de dados que contém áudio ruidoso com palavras foneticamente semelhantes ou idênticas a outras, como know and go, shoe and two, cry and high ou right and write. Nesse caso, os rotuladores podem rotular essas amostras de forma inconsistente.

Para manter um alto grau de correção na rotulagem, uma abordagem comum é usar a votação por maioria, na qual a mesma amostra de dados é fornecida a vários trabalhadores e seus resultados são agregados. Esse método e suas variações mais sofisticadas estão descritos na postagem do

blog Use a sabedoria das multidões com o Amazon SageMaker AI Ground Truth para anotar dados com mais precisão no blog do AWS Machine Learning.

Divisões e vazamento de dados

O vazamento de dados ocorre quando seu modelo obtém dados durante a inferência — no momento em que o modelo está em produção e recebendo solicitações de previsão — aos quais ele não deveria ter acesso, como amostras de dados que foram usadas para treinamento ou informações que não estarão disponíveis quando o modelo for implantado na produção.

Se seu modelo for testado inadvertidamente em dados de treinamento, o vazamento de dados poderá causar sobreajuste. O ajuste excessivo significa que seu modelo não se generaliza bem para dados invisíveis. Esta seção fornece as melhores práticas para evitar vazamento e sobreajuste de dados.

Divida seus dados em pelo menos três conjuntos

Uma fonte comum de vazamento de dados é dividir (dividir) seus dados de forma inadequada durante o treinamento. Por exemplo, o cientista de dados pode ter treinado, consciente ou inconscientemente, o modelo com base nos dados que foram usados para testes. Em tais situações, você pode observar métricas de sucesso muito altas causadas pelo ajuste excessivo. Para resolver esse problema, você deve dividir os dados em pelo menos três conjuntos: `trainingvalidation`, `testing` e.

Ao dividir seus dados dessa forma, você pode usar o `validation` conjunto para escolher e ajustar os parâmetros usados para controlar o processo de aprendizado (hiperparâmetros). Quando você tiver alcançado o resultado desejado ou atingido um patamar de melhoria, faça uma avaliação no `testing` conjunto. As métricas de desempenho do `testing` conjunto devem ser semelhantes às métricas dos outros conjuntos. Isso indica que não há incompatibilidade de distribuição entre os conjuntos, e espera-se que seu modelo se generalize bem na produção.

Use um algoritmo de divisão estratificada

Ao dividir seus dados em `trainingvalidation`, e `testing` para pequenos conjuntos de dados, ou quando você trabalha com dados altamente desequilibrados, certifique-se de usar um algoritmo de divisão estratificada. A estratificação garante que cada divisão contenha aproximadamente o mesmo número ou distribuição de classes para cada divisão. A biblioteca de ML scikit-learn já implementa a estratificação, assim como o Apache Spark.

Para o tamanho da amostra, certifique-se de que os conjuntos de validação e teste tenham dados suficientes para avaliação, para que você possa chegar a conclusões estatisticamente significativas. Por exemplo, um tamanho de divisão comum para conjuntos de dados relativamente pequenos (menos de 1 milhão de amostras) é 70%, 15% e 15%`training`, para`validation`, e. `testing` Para conjuntos de dados muito grandes (mais de 1 milhão de amostras), você pode usar 90%, 5% e 5% para maximizar os dados de treinamento disponíveis.

Em alguns casos de uso, é útil dividir os dados em conjuntos adicionais, pois os dados de produção podem ter sofrido mudanças radicais e repentinas na distribuição durante o período em que estavam sendo coletados. Por exemplo, considere um processo de coleta de dados para criar um modelo de previsão de demanda para itens de mercearia. Se a equipe de ciência de dados coletasse os `training` dados durante 2019 e os `testing` dados de janeiro de 2020 a março de 2020, um modelo provavelmente teria uma boa pontuação no `testing` conjunto. No entanto, quando o modelo é implantado na produção, o padrão de consumo de determinados itens já teria mudado significativamente devido à pandemia de COVID-19, e o modelo geraria resultados ruins. Nesse cenário, faria sentido adicionar outro conjunto (por exemplo,`recent_testing`) como uma salvaguarda adicional para a aprovação do modelo. Essa adição pode impedir que você aprove um modelo para produção que teria um desempenho instantaneamente ruim devido à incompatibilidade de distribuição.

Em alguns casos, talvez você queira criar `testing` conjuntos adicionais `validation` ou que incluam tipos específicos de amostras, como dados associados a populações minoritárias. É importante corrigir essas amostras de dados, mas podem não estar bem representadas no conjunto de dados geral. Esses subconjuntos de dados são chamados de fatias.

Considere o caso de um modelo de ML para análise de crédito que foi treinado em dados de um país inteiro e foi balanceado para contabilizar igualmente todo o domínio da variável-alvo. Além disso, considere que esse modelo pode ter um `City` recurso. Se o banco que usa esse modelo expandir seus negócios para uma cidade específica, talvez esteja interessado em saber como o modelo funciona nessa região. Portanto, um pipeline de aprovação não deve apenas avaliar a qualidade do modelo com base nos dados de teste de todo o país, mas também deve avaliar os dados de teste de uma determinada área da cidade.

Quando os cientistas de dados trabalham em um novo modelo, eles podem avaliar facilmente as capacidades do modelo e considerar os casos extremos integrando fatias sub-representadas no estágio de validação do modelo.

Considere amostras duplicadas ao fazer divisões aleatórias

Outra fonte menos comum de vazamento está em conjuntos de dados que podem conter muitas amostras duplicadas. Nesse caso, mesmo se você dividir os dados em subconjuntos, subconjuntos diferentes podem ter amostras em comum. Dependendo do número de duplicatas, o sobreajuste pode ser confundido com generalização.

Considere recursos que podem não estar disponíveis ao receber inferências na produção

O vazamento de dados também acontece quando os modelos são treinados com recursos que não estão disponíveis na produção, no instante em que as inferências são invocadas. Como os modelos geralmente são criados com base em dados históricos, esses dados podem ser enriquecidos com colunas ou valores adicionais que não estavam presentes em algum momento. Considere o caso de um modelo de aprovação de crédito que tenha um recurso que monitora quantos empréstimos um cliente fez com o banco nos últimos seis meses. Existe o risco de vazamento de dados se esse modelo for implantado e usado para aprovação de crédito para um novo cliente que não tenha um histórico de seis meses com o banco.

A Amazon SageMaker AI Feature Store ajuda a resolver esse problema. Você pode testar seus modelos com mais precisão com o uso de consultas de viagem no tempo, que podem ser usadas para visualizar dados em momentos específicos.

Loja de recursos

Usar o SageMaker AI Feature Store aumenta a produtividade da equipe, pois separa os limites dos componentes (por exemplo, armazenamento versus uso). Ele também fornece reutilização de recursos em diferentes equipes de ciência de dados em sua organização.

Use consultas de viagem no tempo

Os recursos de viagem no tempo na Feature Store ajudam a reproduzir construções de modelos e apoiam práticas de governança mais fortes. Isso pode ser útil quando uma organização deseja avaliar a linhagem de dados, da mesma forma que ferramentas de controle de versão, como o Git, avaliam o código. As consultas de viagem no tempo também ajudam as organizações a fornecer dados precisos para verificações de conformidade. Para obter mais informações, consulte Entendendo os principais recursos da Amazon SageMaker AI Feature Store no blog do AWS Machine Learning.

Usar funções do IAM

O Feature Store também ajuda a melhorar a segurança sem afetar a produtividade e a inovação da equipe. Você pode usar funções AWS Identity and Access Management (IAM) para conceder ou restringir o acesso granular a recursos específicos para usuários ou grupos específicos.

Por exemplo, a política a seguir restringe o acesso a um recurso confidencial na Feature Store.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "VisualEditor0",
            "Effect": "Deny",
            "Action": "*",
            "Resource": "arn:aws:s3:::amzn-s3-demo-bucket--usw2-az1--x-s3/12345678910/
sagemaker/us-east-2/offline-store/doctor-appointments"
        }
    ]
}
```

Para obter mais informações sobre segurança e criptografia de dados usando o Feature Store, consulte Segurança e controle de acesso na documentação de SageMaker IA.

Use testes unitários

Quando os cientistas de dados criam modelos com base em alguns dados, geralmente fazem suposições sobre a distribuição dos dados ou realizam uma análise completa para entender completamente as propriedades dos dados. Quando esses modelos são implantados, eles acabam ficando obsoletos. Quando o conjunto de dados fica desatualizado, cientistas de dados, engenheiros de ML e (em alguns casos) sistemas automatizados retreinam o modelo com novos dados obtidos em uma loja on-line ou off-line.

No entanto, a distribuição desses novos dados pode ter mudado, o que pode afetar o desempenho do algoritmo atual. Uma forma automatizada de verificar esses tipos de problemas é emprestar o conceito de teste unitário da engenharia de software. Coisas comuns a serem testadas incluem a porcentagem de valores ausentes, a cardinalidade das variáveis categóricas e se as colunas de valores reais aderem a alguma distribuição esperada usando uma estrutura como estatísticas de teste de hipóteses (teste t). Talvez você também queira validar o esquema de dados para garantir que ele não tenha sido alterado e não gere recursos de entrada inválidos silenciosamente.

O teste unitário exige a compreensão dos dados e de seu domínio para que você possa planejar as afirmações exatas a serem executadas como parte do projeto de ML. Para obter mais informações, consulte Testando a qualidade dos dados em grande PyDeequ escala com o blog AWS Big Data.

Treinamento

MLOps está preocupado com a operacionalização do ciclo de vida do ML. Portanto, deve facilitar o trabalho dos cientistas e engenheiros de dados na criação de modelos pragmáticos que atendam às necessidades de negócios e funcionem bem a longo prazo, sem incorrer em dívidas técnicas.

Siga as melhores práticas nesta seção para ajudar a enfrentar os desafios do treinamento de modelos.

Tópicos

Crie um modelo de linha de base

Quando os profissionais enfrentam um problema comercial com uma solução de ML, normalmente sua primeira inclinação é usar o state-of-the-art algoritmo. Essa prática é arriscada, pois é provável que o state-of-the-art algoritmo não tenha sido testado pelo tempo. Além disso, o state-of-the-art algoritmo geralmente é mais complexo e não é bem compreendido, portanto, pode resultar em apenas melhorias marginais em relação a modelos alternativos mais simples. Uma prática melhor é criar um modelo básico que seja relativamente rápido de validar e implantar e que possa conquistar a confiança das partes interessadas do projeto.

Ao criar uma linha de base, recomendamos que você avalie seu desempenho métrico sempre que possível. Compare o desempenho do modelo básico com outros sistemas automatizados ou manuais para garantir seu sucesso e garantir que a implementação do modelo ou o projeto possam ser entregues a médio e longo prazo.

O modelo básico deve ser validado ainda mais com engenheiros de ML para confirmar se o modelo pode fornecer os requisitos não funcionais que foram estabelecidos para o projeto, como tempo de inferência, com que frequência se espera que os dados mudem de distribuição, se o modelo pode ser facilmente retreinado nesses casos e como ele será implantado, o que afetará o custo da

solução. Obtenha pontos de vista multidisciplinares sobre essas questões para aumentar a chance de você desenvolver um modelo bem-sucedido e duradouro.

Os cientistas de dados podem estar inclinados a adicionar o máximo de recursos possível a um modelo básico. Embora isso aumente a capacidade de um modelo de prever o resultado desejado, alguns desses recursos podem gerar apenas melhorias métricas incrementais. Muitos recursos, especialmente aqueles altamente correlacionados, podem ser redundantes. Adicionar muitos recursos aumenta os custos, pois exige mais recursos computacionais e ajustes. Muitos recursos também afetam day-to-day as operações do modelo, porque a variação de dados se torna mais provável ou ocorre mais rapidamente.

Considere um modelo no qual dois recursos de entrada são altamente correlacionados, mas apenas um recurso tem causalidade. Por exemplo, um modelo que prevê se um empréstimo será inadimplente pode ter características de entrada, como idade e renda do cliente, que podem estar altamente correlacionadas, mas somente a renda deve ser usada para conceder ou negar um empréstimo. Um modelo que foi treinado com esses dois recursos pode estar contando com o recurso que não tem causalidade, como idade, para gerar a saída da previsão. Se, depois de entrar em produção, o modelo receber solicitações de inferência para clientes maiores ou menores do que a idade média incluída no conjunto de treinamento, ele poderá começar a ter um desempenho ruim.

Além disso, cada recurso individual pode sofrer uma mudança na distribuição durante a produção e fazer com que o modelo se comporte de forma inesperada. Por esses motivos, quanto mais características um modelo tem, mais frágil ele é em relação à deriva e à obsolescência.

Os cientistas de dados devem usar medidas de correlação e valores de Shapley para avaliar quais características agregam valor suficiente à previsão e devem ser mantidas. Ter modelos tão complexos aumenta a chance de um ciclo de feedback, no qual o modelo muda o ambiente para o qual foi modelado. Um exemplo é um sistema de recomendação no qual o comportamento do consumidor pode mudar devido às recomendações de um modelo. Os ciclos de feedback que atuam em vários modelos são menos comuns. Por exemplo, considere um sistema de recomendação que recomenda filmes e outro sistema que recomenda livros. Se os dois modelos visarem o mesmo conjunto de consumidores, eles afetarão um ao outro.

Para cada modelo que você desenvolver, considere quais fatores podem contribuir para essa dinâmica, para que você saiba quais métricas monitorar na produção.

Use uma abordagem centrada em dados e análise de erros

Se você usa um modelo simples, sua equipe de ML pode se concentrar em melhorar os dados em si e adotar uma abordagem centrada em dados em vez de uma abordagem centrada em modelos. Se seu projeto usa dados não estruturados, como imagens, texto, áudio e outros formatos que podem ser avaliados por humanos (em comparação com dados estruturados, que podem ser mais difíceis de mapear em um rótulo de forma eficiente), uma boa prática para obter um melhor desempenho do modelo é realizar a análise de erros.

A análise de erros envolve avaliar um modelo em um conjunto de validação e verificar os erros mais comuns. Isso ajuda a identificar grupos potenciais de amostras de dados semelhantes que o modelo pode estar tendo dificuldades em corrigir. Para realizar a análise de erros, você pode listar inferências que tiveram maiores erros de predição ou classificar erros nos quais uma amostra de uma classe foi prevista como sendo de outra classe, por exemplo.

Arquitete seu modelo para uma iteração rápida

Quando os cientistas de dados seguem as melhores práticas, eles podem experimentar um novo algoritmo ou combinar diferentes recursos com facilidade e rapidez durante a prova de conceito ou até mesmo a reciclagem. Essa experimentação contribui para o sucesso na produção. Uma boa prática é desenvolver o modelo básico, empregando algoritmos um pouco mais complexos e adicionando novos recursos de forma iterativa, enquanto monitora o desempenho no conjunto de treinamento e validação para comparar o comportamento real com o comportamento esperado. Essa estrutura de treinamento pode fornecer um equilíbrio ideal no poder de previsão e ajudar a manter os modelos o mais simples possível com uma menor pegada de débito técnico.

Para uma iteração rápida, os cientistas de dados devem trocar diferentes implementações de modelos para determinar o melhor modelo a ser usado para dados específicos. Se você tem uma equipe grande, um prazo curto e outras logísticas relacionadas ao gerenciamento de projetos, a iteração rápida pode ser difícil sem um método estabelecido.

Em engenharia de software, o princípio de substituição de Liskov é um mecanismo para arquitetar interações entre componentes de software. Esse princípio afirma que você deve ser capaz de substituir uma implementação de uma interface por outra implementação sem interromper o aplicativo cliente ou a implementação. Ao escrever código de treinamento para seu sistema de ML, você pode empregar esse princípio para estabelecer limites e encapsular o código, para poder substituir o algoritmo com facilidade e experimentar novos algoritmos com mais eficiência.

Por exemplo, no código a seguir, você pode adicionar novos experimentos simplesmente adicionando uma nova implementação de classe.

```
from abc import ABC, abstractmethod

from pandas import DataFrame


class ExperimentRunner(object):

    def __init__(self, *experiments):
        self.experiments = experiments

    def run(self, df: DataFrame) -> None:
        for experiment in self.experiments:
            result = experiment.run(df)
            print(f'Experiment "{experiment.name}" gave result {result}')


class Experiment(ABC):

    @abstractmethod
    def run(self, df: DataFrame) -> float:
        pass

    @property
    @abstractmethod
    def name(self) -> str:
        pass


class Experiment1(Experiment):

    def run(self, df: DataFrame) -> float:
        print('performing experiment 1')
        return 0

    def name(self) -> str:
        return 'experiment 1'


class Experiment2(Experiment):
```

```
    def run(self, df: DataFrame) -> float:
        print('performing experiment 2')
        return 0

    def name(self) -> str:
        return 'experiment 2'


class Experiment3(Experiment):

    def run(self, df: DataFrame) -> float:
        print('performing experiment 3')
        return 0

    def name(self) -> str:
        return 'experiment 3'


if __name__ == '__main__':
    runner = ExperimentRunner(*[
        Experiment1(),
        Experiment2(),
        Experiment3()
    ])
    df = ...
    runner.run(df)
```

Acompanhe seus experimentos de ML

Quando você trabalha com um grande número de experimentos, é importante avaliar se as melhorias observadas são produto de mudanças implementadas ou do acaso. Você pode usar o Amazon SageMaker AI Experiments para criar facilmente experimentos e associar metadados a eles para rastreamento, comparação e avaliação.

Reduzir a aleatoriedade do processo de criação do modelo é útil para depuração, solução de problemas e melhoria da governança, pois você pode prever a inferência do modelo de saída com mais certeza, considerando o mesmo código e dados.

Muitas vezes, não é possível tornar um código de treinamento totalmente reproduzível, devido à inicialização aleatória do peso, à sincronicidade computacional paralela, às complexidades internas da GPU e a fatores não determinísticos semelhantes. No entanto, definir corretamente as sementes

aleatórias, para garantir que cada corrida de treinamento comece do mesmo ponto e se comporte de forma semelhante, melhora significativamente a previsibilidade dos resultados.

Solucionar problemas de tarefas de treinamento

Em alguns casos, pode ser difícil para os cientistas de dados ajustar até mesmo um modelo básico muito simples. Nesse caso, eles podem decidir que precisam de um algoritmo que possa se ajustar melhor a funções complexas. Um bom teste é usar a linha de base de uma parte muito pequena do conjunto de dados (por exemplo, cerca de 10 amostras) para garantir que o algoritmo se ajuste demais a essa amostra. Isso ajuda a descartar problemas de dados ou código.

Outra ferramenta útil para depurar cenários complexos é o Amazon SageMaker AI Debugger, que pode capturar problemas relacionados à correção algorítmica e à infraestrutura, como o uso ideal da computação.

Implantação

Na engenharia de software, colocar código em produção exige a devida diligência, pois o código pode se comportar de forma inesperada, um comportamento imprevisto do usuário pode danificar o software e casos extremos inesperados podem ser encontrados. Os engenheiros e DevOps engenheiros de software geralmente empregam testes unitários e estratégias de reversão para mitigar esses riscos. Com o ML, colocar modelos em produção exige ainda mais planejamento, porque se espera que o ambiente real mude e, em muitas ocasiões, os modelos são validados em métricas que são proxies das métricas reais de negócios que eles estão tentando melhorar.

Siga as melhores práticas nesta seção para ajudar a enfrentar esses desafios.

Tópicos

Automatize o ciclo de implantação

O processo de treinamento e implantação deve ser totalmente automatizado para evitar erros humanos e garantir que as verificações de compilação sejam executadas de forma consistente. Os usuários não devem ter permissões de acesso de gravação ao ambiente de produção.

O Amazon SageMaker AI Pipelines e AWS CodePipelinea ajuda a criar um CI/CD pipelines for ML projects. One of the advantages of using a CI/CD pipeline permitem que todo código usado para ingerir dados, treinar um modelo e realizar monitoramento possa ser controlado por versão usando uma ferramenta como o Git. Às vezes, você precisa retreinar um modelo usando o mesmo algoritmo e hiperparâmetros, mas com dados diferentes. A única maneira de verificar se você está usando a versão correta do algoritmo é usar o controle de origem e as tags. Você pode usar os modelos de projeto padrão fornecidos pela SageMaker IA como ponto de partida para sua MLOps prática.

Ao criar pipelines de CI/CD para implantar seu modelo, certifique-se de marcar seus artefatos de construção com um identificador de compilação, versão ou confirmação do código e versão de dados. Essa prática ajuda você a solucionar quaisquer problemas de implantação. Às vezes, a marcação também é necessária para modelos que fazem previsões em campos altamente regulamentados. A capacidade de retroceder e identificar os dados, códigos, compilações,

verificações e aprovações exatos associados a um modelo de ML pode ajudar a melhorar significativamente a governança.

Parte do trabalho do pipeline de CI/CD é realizar testes sobre o que ele está construindo. Embora se espere que os testes de unidade de dados ocorram antes que os dados sejam ingeridos por uma feature store, o pipeline ainda é responsável por realizar testes na entrada e na saída de um determinado modelo e por verificar as principais métricas. Um exemplo dessa verificação é validar um novo modelo em um conjunto fixo de validação e confirmar que seu desempenho é semelhante ao modelo anterior usando um limite estabelecido. Se o desempenho for significativamente menor do que o esperado, a construção deve falhar e o modelo não deve entrar em produção.

O uso extensivo de pipelines de CI/CD também oferece suporte a pull requests, que ajudam a evitar erros humanos. Quando você usa pull requests, cada alteração de código deve ser revisada e aprovada por pelo menos um outro membro da equipe antes de entrar em produção. As pull requests também são úteis para identificar códigos que não cumprem as regras de negócios e para difundir o conhecimento dentro da equipe.

Escolha uma estratégia de implantação

MLOps as estratégias de implantação incluem blue/green, canary, shadow, and A/B testes.

Azul/verde

Blue/green deployments are very common in software development. In this mode, two systems are kept running during development: blue is the old environment (in this case, the model that is being replaced) and green is the newly released model that is going to production. Changes can easily be rolled back with minimum downtime, because the old system is kept alive. For more in-depth information about blue/greenimplantações no contexto de SageMaker, consulte a postagem do blog Implantação e monitoramento seguros de endpoints de SageMaker IA da Amazon com AWS CodePipeline e AWS CodeDeploy no blog do AWS Machine Learning.

Canário

As implantações do Canary são semelhantes às blue/green deployments in that both keep two models running together. However, in canary deployments, the new model is rolled out to users incrementally, until all traffic eventually shifts over to the new model. As in blue/green implantações; o risco é mitigado porque o novo modelo (e potencialmente defeituoso) é monitorado de perto durante

a implantação inicial e pode ser revertido em caso de problemas. Na SageMaker IA, você pode especificar a distribuição inicial do tráfego usando a InitialVariantWeightAPI.

Shadow

Você pode usar implantações paralelas para colocar um modelo em produção com segurança. Nesse modo, o novo modelo funciona junto com um modelo ou processo de negócios mais antigo e realiza inferências sem influenciar nenhuma decisão. Esse modo pode ser útil como uma verificação final ou um experimento de maior fidelidade antes de você promover o modelo para produção.

O modo sombra é útil quando você não precisa de nenhum feedback de inferência do usuário. Você pode avaliar a qualidade das previsões realizando análises de erros e comparando o novo modelo com o modelo antigo, além de monitorar a distribuição de saída para verificar se está conforme o esperado. Para ver como fazer a implantação paralela com SageMaker IA, consulte a postagem do blog Implante modelos de ML paralelos na Amazon SageMaker AI no blog do AWS Machine Learning.

Testes A/B

Quando os profissionais de ML desenvolvem modelos em seus ambientes, as métricas para as quais eles otimizam geralmente são proxies das métricas de negócios que realmente importam. Isso torna difícil dizer com certeza se um novo modelo realmente melhorará os resultados comerciais, como receita e taxa de cliques, e reduzirá o número de reclamações de usuários.

Considere o caso de um site de comércio eletrônico em que o objetivo comercial é vender o maior número possível de produtos. A equipe de avaliação sabe que as vendas e a satisfação do cliente se correlacionam diretamente com avaliações informativas e precisas. Um membro da equipe pode propor um novo algoritmo de classificação de avaliações para melhorar as vendas. Ao usar o teste A/B, eles poderiam implantar os algoritmos antigos e novos em grupos de usuários diferentes, mas semelhantes, e monitorar os resultados para ver se os usuários que receberam previsões do modelo mais novo têm maior probabilidade de fazer compras.

O teste A/B também ajuda a avaliar o impacto comercial da obsolescência e do desvio do modelo. As equipes podem colocar novos modelos em produção com alguma recorrência, realizar testes A/B com cada modelo e criar um gráfico de idade versus desempenho. Isso ajudaria a equipe a entender a volatilidade do desvio de dados em seus dados de produção.

Para obter mais informações sobre como realizar testes A/B com SageMaker IA, consulte a postagem do blog Teste A/B de modelos ML em produção usando o Amazon SageMaker AI no blog do AWS Machine Learning.

Considere seus requisitos de inferência

Com a SageMaker IA, você pode escolher a infraestrutura subjacente para implantar seu modelo de maneiras diferentes. Esses recursos de invocação de inferência oferecem suporte a diferentes casos de uso e perfis de custo. Suas opções incluem inferência em tempo real, inferência assíncrona e transformação em lote, conforme discutido nas seções a seguir.

Inferência em tempo real

A inferência em tempo real é ideal para cargas de trabalho de inferência em que você tem requisitos em tempo real, interativos e de baixa latência. Você pode implantar seu modelo em serviços de hospedagem de SageMaker IA e obter um endpoint que pode ser usado para inferência. Esses endpoints são totalmente gerenciados, oferecem suporte à escalabilidade automática (consulte Escalar automaticamente os modelos de SageMaker IA da Amazon) e podem ser implantados em várias zonas de disponibilidade.

Se você tem um modelo de aprendizado profundo criado com o Apache MXNet, PyTorch TensorFlow, ou também pode usar o Amazon SageMaker AI Elastic Inference (EI). Com o EI, você pode anexar frações GPUs a qualquer instância de SageMaker IA para acelerar a inferência. Você pode selecionar a instância do cliente para executar seu aplicativo e anexar um acelerador de EI para usar a quantidade correta de aceleração de GPU para suas necessidades de inferência.

Outra opção é usar endpoints de vários modelos, que fornecem uma solução escalável e econômica para a implantação de um grande número de modelos. Esses endpoints usam um contêiner de serviço compartilhado que está habilitado para hospedar vários modelos. Os endpoints multimodelo reduzem os custos de hospedagem melhorando a utilização do endpoint em comparação com o uso de endpoints de modelo único. Eles também reduzem a sobrecarga de implantação, porque a SageMaker IA gerencia o carregamento de modelos na memória e a escalabilidade deles com base nos padrões de tráfego.

Para obter mais práticas recomendadas para implantar modelos de ML em SageMaker IA, consulte Melhores práticas de implantação na documentação de SageMaker IA.

Inferência assíncrona

O Amazon SageMaker AI Asynchronous Inference é um recurso de SageMaker IA que enfileira as solicitações recebidas e as processa de forma assíncrona. Essa opção é ideal para solicitações com grandes cargas de até 1 GB, tempos de processamento longos e requisitos de latência quase em

tempo real. A inferência assíncrona permite que você economize custos escalando automaticamente a contagem de instâncias para zero quando não há solicitações para processar, então você paga somente quando seu endpoint está processando solicitações.

Transformação em lote

Use a transformação em lote quando quiser fazer o seguinte:

- Pré-processar os conjuntos de dados para remover ruído ou desvio que interfira no treinamento ou na inferência do conjunto de dados.
- Obter inferências de conjuntos de dados grandes.
- Executar inferência quando não for necessário um endpoint persistente.
- Associar registros de entrada a inferências para auxiliar na interpretação de resultados.

Monitoramento

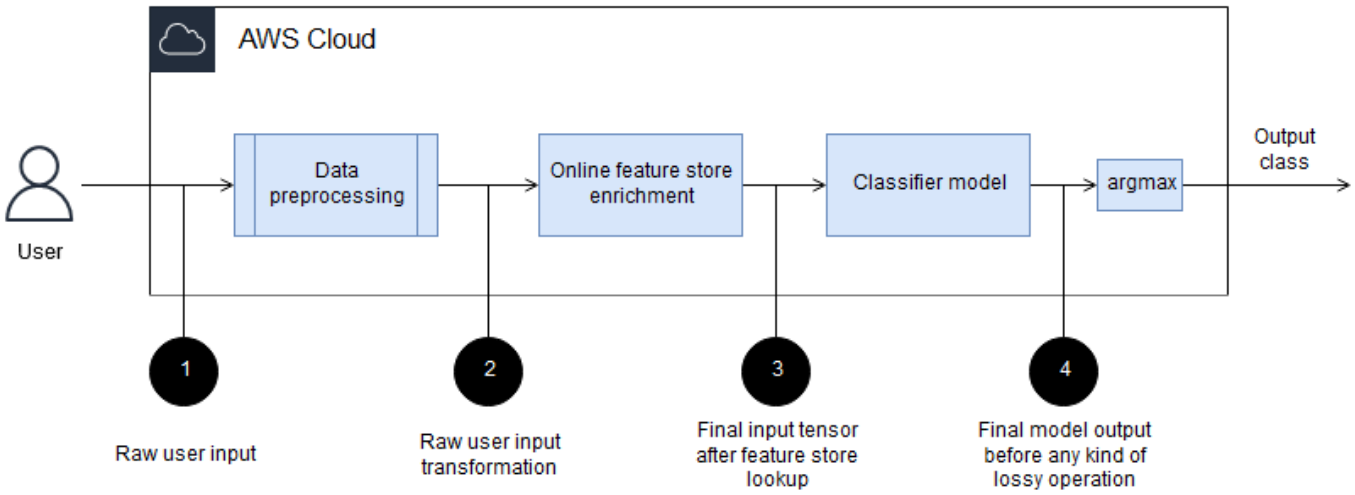
Quando os modelos já estiverem em produção e agregando valor comercial, execute verificações contínuas para identificar quando os modelos devem ser retreinados ou executados.

Sua equipe de monitoramento deve se comportar de forma proativa, não reativa, para entender melhor o comportamento dos dados do ambiente e identificar a frequência, a taxa e a abrupção dos desvios de dados. A equipe deve identificar novos casos extremos nos dados que possam estar sub-representados no conjunto de treinamento, no conjunto de validação e em outras fatias de casos extremos. Eles devem armazenar métricas de qualidade de serviço (QoS), usar alarmes para agir imediatamente quando surgir um problema e definir uma estratégia para ingerir e alterar os conjuntos de dados atuais. Essas práticas começam registrando solicitações e respostas para o modelo, a fim de fornecer uma referência para solução de problemas ou informações adicionais.

Idealmente, as transformações de dados devem ser registradas em alguns estágios importantes durante o processamento:

- Antes de qualquer tipo de pré-processamento
- Depois de qualquer tipo de enriquecimento da feature store
- Afinal, os principais estágios de um modelo
- Antes de qualquer tipo de função com perdas na saída do modelo, como `argmax`

O diagrama a seguir ilustra esses estágios.



Você pode usar o SageMaker AI Model Monitor para capturar automaticamente dados de entrada e saída e armazená-los no Amazon Simple Storage Service (Amazon S3). Você pode implementar outros tipos de registro intermediário adicionando registros a um contêiner de serviço personalizado.

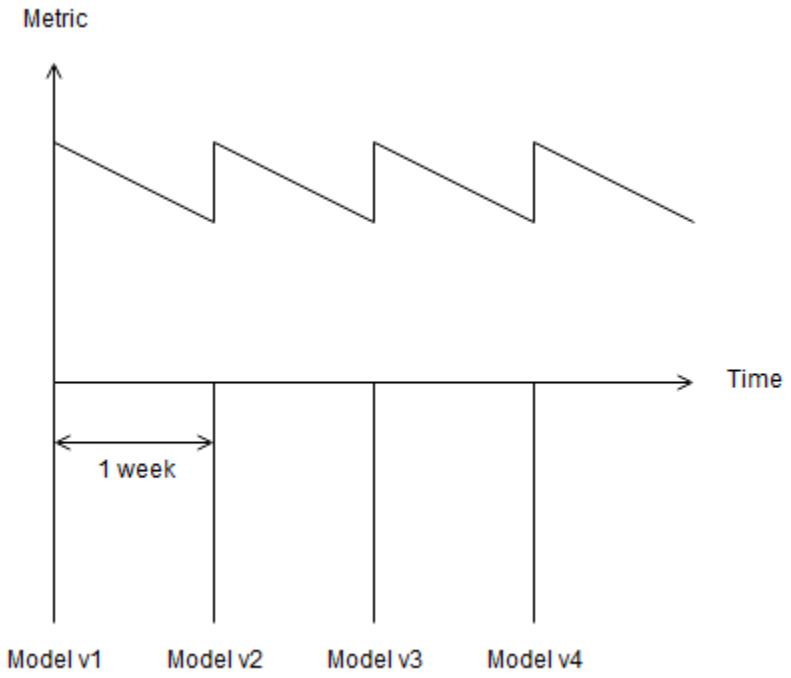
Depois de registrar os dados dos modelos, você pode monitorar o desvio da distribuição. Em alguns casos, você pode obter a verdade básica (dados rotulados corretamente) logo após a inferência. Um exemplo comum disso é um modelo que prevê os anúncios mais relevantes a serem exibidos para um usuário. Assim que o usuário sair da página, você poderá determinar se ele clicou no anúncio. Se o usuário clicou no anúncio, você pode registrar essas informações. Neste exemplo simples, você pode quantificar facilmente o sucesso do seu modelo usando uma métrica, como precisão ou F1, que pode ser medida tanto no treinamento quanto na implantação. Para obter mais informações sobre esses cenários nos quais você rotulou dados, consulte Monitorar a qualidade do modelo na documentação de SageMaker IA. No entanto, esses cenários simples não são frequentes, porque os modelos geralmente são projetados para otimizar métricas matematicamente convenientes que são apenas indicadores dos resultados comerciais reais. Nesses casos, a melhor prática é monitorar o resultado comercial quando um modelo é implantado na produção.

Considere o caso de um modelo de classificação de avaliações. Se o resultado comercial definido do modelo de ML for exibir as avaliações mais relevantes e úteis na parte superior da página da Web, você poderá medir o sucesso do modelo adicionando um botão como “Isso foi útil?” para cada avaliação. Medir a taxa de cliques desse botão pode ser uma medida de resultados comerciais que ajuda você a medir o desempenho do seu modelo na produção.

Para monitorar o desvio das etiquetas de entrada ou saída na SageMaker IA, você pode usar os recursos de qualidade de dados do SageMaker AI Model Monitor, que monitoram tanto a entrada quanto a saída. Você também pode implementar sua própria lógica para o SageMaker AI Model Monitor criando um contêiner personalizado.

Monitorar os dados que um modelo recebe tanto no tempo de desenvolvimento quanto no tempo de execução é fundamental. Os engenheiros devem monitorar os dados não apenas em busca de alterações no esquema, mas também em busca de incompatibilidades de distribuição. Detectar alterações no esquema é mais fácil e pode ser implementado por meio de um conjunto de regras, mas a incompatibilidade de distribuição geralmente é mais complicada, especialmente porque exige que você defina um limite para quantificar quando acionar um alarme. Nos casos em que a distribuição monitorada é conhecida, geralmente a maneira mais fácil é monitorar os parâmetros da distribuição. No caso de uma distribuição normal, essa seria a média e o desvio padrão. Outras métricas importantes, como a porcentagem de valores faltantes, valores máximos e valores mínimos, também são úteis.

Você também pode criar trabalhos de monitoramento contínuo que coletem amostras de dados de treinamento e dados de inferência e comparem suas distribuições. Você pode criar essas tarefas para entrada e saída do modelo e representar graficamente os dados em relação ao tempo para visualizar qualquer desvio repentino ou gradual. Isso é ilustrado no gráfico a seguir.



Para entender melhor o perfil de variação dos dados, como a frequência com que a distribuição de dados muda significativamente, em que taxa ou com que rapidez, recomendamos que você implante continuamente novas versões do modelo e monitore seu desempenho. Por exemplo, se sua equipe implanta um novo modelo toda semana e observa que o desempenho do modelo melhora significativamente a cada vez, eles podem determinar que devem entregar novos modelos em menos de uma semana, no mínimo.

Próximas etapas e recursos

Este guia explica algumas considerações ao planejar o ciclo de vida dos modelos de aprendizado de máquina que você deseja levar à produção. Ele discute os desafios e as melhores práticas em quatro áreas — dados, treinamento, implantação e monitoramento — e inclui recursos adicionais relevantes.

AWS fornece o Well-Architected Framework, que ajuda os arquitetos de nuvem a criar infraestruturas seguras, de alto desempenho, resilientes e eficientes para uma variedade de aplicativos, cargas de trabalho e domínios de tecnologia. Para ler mais, consulte o Machine Learning Lens oferecido pela AWS Well-Architected.

Recursos

Documentação da Amazon SageMaker AI

- Loja de recursos de SageMaker IA da Amazon
- Segurança e controle de acesso da Feature Store
- Valores de Shapley
- Depurador de SageMaker IA da Amazon
- Pipelines de SageMaker IA da Amazon
- Modelos de projeto padrão do Amazon SageMaker AI
- SageMaker Inferência de IA em tempo real
- Dimensione automaticamente os modelos de SageMaker IA da Amazon
- Inferência assíncrona de SageMaker IA da Amazon
- SageMaker Monitor de modelo AI

AWS ferramentas para desenvolvedores

- AWS CodePipeline

AWS postagens no blog

- Entendendo os principais recursos da Amazon SageMaker AI Feature Store
- Testando a qualidade dos dados em grande escala com PyDeequ

- Experiências de SageMaker IA da Amazon
- Implantação e monitoramento seguros de SageMaker endpoints da Amazon com e CodePipeline AWS CodeDeploy
- Implemente modelos de ML paralelos na Amazon SageMaker AI
- Teste A/B de modelos de ML em produção usando Amazon AI SageMaker

Histórico do documento

A tabela a seguir descreve alterações significativas feitas neste guia. Se desejar receber notificações sobre futuras atualizações, inscreva-se em um feed RSS.

Alteração	Descrição	Data
Publicação inicial	—	20 de dezembro de 2021

AWS Glossário de orientação prescritiva

A seguir estão os termos comumente usados em estratégias, guias e padrões fornecidos pela Orientação AWS Prescritiva. Para sugerir entradas, use o link Fornecer feedback no final do glossário.

Números

7 Rs

Sete estratégias comuns de migração para mover aplicações para a nuvem. Essas estratégias baseiam-se nos 5 Rs identificados pela Gartner em 2011 e consistem em:

- Refatorar/rearquitetar: mova uma aplicação e modifique sua arquitetura aproveitando ao máximo os recursos nativos de nuvem para melhorar a agilidade, a performance e a escalabilidade. Isso normalmente envolve a portabilidade do sistema operacional e do banco de dados. Exemplo: migrar seu banco de dados Oracle on-premises para o Amazon Aurora Edição Compatível com PostgreSQL.
- Redefinir a plataforma (mover e redefinir [mover e redefinir (lift-and-reshape]): mova uma aplicação para a nuvem e introduza algum nível de otimização a fim de aproveitar os recursos da nuvem. Exemplo: migrar seu banco de dados Oracle on-premises para o Amazon Relational Database Service (Amazon RDS) para Oracle na Nuvem AWS.
- Recomprar (drop and shop): mude para um produto diferente, normalmente migrando de uma licença tradicional para um modelo SaaS. Exemplo: migrar seu sistema de gerenciamento de relacionamento com o cliente (CRM) para o Salesforce.com.
- Redefinir a hospedagem (mover sem alterações [lift-and-shift])mover uma aplicação para a nuvem sem fazer nenhuma alteração a fim de aproveitar os recursos da nuvem. Exemplo: migrar seu banco de dados Oracle on-premises para o Oracle em uma instância do EC2 na Nuvem AWS.
- Realocar (mover o hipervisor sem alterações [hypervisor-level lift-and-shift]): mover a infraestrutura para a nuvem sem comprar novo hardware, reescrever aplicações ou modificar suas operações existentes. Você migra servidores de uma plataforma on-premises para um serviço de nuvem para a mesma plataforma. Exemplo: Migrar um Microsoft Hyper-V aplicativo para o. AWS
- Reter (revisitar): mantenha as aplicações em seu ambiente de origem. Isso pode incluir aplicações que exigem grande refatoração, e você deseja adiar esse trabalho para um

momento posterior, e aplicações antigas que você deseja manter porque não há justificativa comercial para migrá-las.

- Retirar: desative ou remova aplicações que não são mais necessárias em seu ambiente de origem.

A

ABAC

Consulte controle de acesso baseado em atributo.

serviços abstraídos

Veja serviços gerenciados.

ACID

Veja atomicidade, consistência, isolamento, durabilidade.

migração ativa-ativa

Um método de migração de banco de dados no qual os bancos de dados de origem e de destino são mantidos em sincronia (por meio de uma ferramenta de replicação bidirecional ou operações de gravação dupla), e ambos os bancos de dados lidam com transações de aplicações conectadas durante a migração. Esse método oferece suporte à migração em lotes pequenos e controlados, em vez de exigir uma substituição única. É mais flexível, mas exige mais trabalho do que a migração ativa-passiva.

migração ativa-passiva

Um método de migração de banco de dados em que os bancos de dados de origem e de destino são mantidos em sincronia, mas somente o banco de dados de origem manipula as transações das aplicações conectadas, enquanto os dados são replicados no banco de dados de destino. O banco de dados de destino não aceita nenhuma transação durante a migração.

AGGREGATE FUNCTION

Uma função SQL que opera em um grupo de linhas e calcula um único valor de retorno para o grupo. Exemplos de funções agregadas incluem `SUM` e `MAX`.

AI

Veja inteligência artificial.

AIOps

Veja operações de inteligência artificial.

anonimização

O processo de excluir permanentemente informações pessoais em um conjunto de dados. A anonimização pode ajudar a proteger a privacidade pessoal. Dados anônimos não são mais considerados dados pessoais.

antipadrões

Uma solução frequentemente usada para um problema recorrente em que a solução é contraproducente, ineficaz ou menos eficaz do que uma alternativa.

controle de aplicações

Uma abordagem de segurança que permite o uso somente de aplicações aprovadas para ajudar a proteger um sistema contra malware.

portfólio de aplicações

Uma coleção de informações detalhadas sobre cada aplicação usada por uma organização, incluindo o custo para criar e manter a aplicação e seu valor comercial. Essas informações são fundamentais para o processo de descoberta e análise de portfólio e ajudam a identificar e priorizar as aplicações a serem migradas, modernizadas e otimizadas.

inteligência artificial (IA)

O campo da ciência da computação que se dedica ao uso de tecnologias de computação para desempenhar funções cognitivas normalmente associadas aos humanos, como aprender, resolver problemas e reconhecer padrões. Para obter mais informações, consulte O que é inteligência artificial?

operações de inteligência artificial (AIOps)

O processo de usar técnicas de machine learning para resolver problemas operacionais, reduzir incidentes operacionais e intervenção humana e aumentar a qualidade do serviço. Para obter mais informações sobre como AIOps é usado na estratégia de AWS migração, consulte o guia de integração de operações.

criptografia assimétrica

Um algoritmo de criptografia que usa um par de chaves, uma chave pública para criptografia e uma chave privada para descriptografia. É possível compartilhar a chave pública porque ela não é usada na descriptografia, mas o acesso à chave privada deve ser altamente restrito.

atomicidade, consistência, isolamento, durabilidade (ACID)

Um conjunto de propriedades de software que garantem a validade dos dados e a confiabilidade operacional de um banco de dados, mesmo no caso de erros, falhas de energia ou outros problemas.

controle de acesso por atributo (ABAC)

A prática de criar permissões minuciosas com base nos atributos do usuário, como departamento, cargo e nome da equipe. Para obter mais informações, consulte ABAC AWS na documentação AWS Identity and Access Management (IAM).

fonte de dados autorizada

Um local onde você armazena a versão principal dos dados, que é considerada a fonte de informações mais confiável. Você pode copiar dados da fonte de dados autorizada para outros locais com o objetivo de processar ou modificar os dados, como anonimizá-los, redigi-los ou pseudonimizá-los.

Zona de disponibilidade

Um local distinto dentro de um Região da AWS que está isolado de falhas em outras zonas de disponibilidade e fornece conectividade de rede barata e de baixa latência a outras zonas de disponibilidade na mesma região.

AWS Estrutura de adoção da nuvem (AWS CAF)

Uma estrutura de diretrizes e melhores práticas AWS para ajudar as organizações a desenvolver um plano eficiente e eficaz para migrar com sucesso para a nuvem. AWS O CAF organiza a orientação em seis áreas de foco chamadas perspectivas: negócios, pessoas, governança, plataforma, segurança e operações. As perspectivas de negócios, pessoas e governança têm como foco habilidades e processos de negócios; as perspectivas de plataforma, segurança e operações concentram-se em habilidades e processos técnicos. Por exemplo, a perspectiva das pessoas tem como alvo as partes interessadas que lidam com recursos humanos (RH), funções de pessoal e gerenciamento de pessoal. Nessa perspectiva, o AWS CAF fornece orientação para desenvolvimento, treinamento e comunicação de pessoas para ajudar a preparar a organização para a adoção bem-sucedida da nuvem. Para obter mais informações, consulte o site da AWS CAF e o whitepaper da AWS CAF.

AWS Estrutura de qualificação da carga de trabalho (AWS WQF)

Uma ferramenta que avalia as cargas de trabalho de migração do banco de dados, recomenda estratégias de migração e fornece estimativas de trabalho. AWS O WQF está incluído com AWS

Schema Conversion Tool ()AWS SCT. Ela analisa esquemas de banco de dados e objetos de código, código de aplicações, dependências e características de performance, além de fornecer relatórios de avaliação.

B

bot malicioso

Um bot destinado a causar disrupção ou danos a indivíduos ou organizações.

BCP

Veja planejamento de continuidade de negócios

gráfico de comportamento

Uma visualização unificada e interativa do comportamento e das interações de recursos ao longo do tempo. É possível usar um gráfico de comportamento com o Amazon Detective para examinar tentativas de login malsucedidas, chamadas de API suspeitas e ações similares. Para obter mais informações, consulte Dados em um gráfico de comportamento na documentação do Detective.

sistema big-endian

Um sistema que armazena o byte mais significativo antes. Veja também endianness.

classificação binária

Um processo que prevê um resultado binário (uma de duas classes possíveis). Por exemplo, seu modelo de ML pode precisar prever problemas como “Este e-mail é ou não é spam?” ou “Este produto é um livro ou um carro?”

filtro de bloom

Uma estrutura de dados probabilística e eficiente em termos de memória que é usada para testar se um elemento é membro de um conjunto.

blue/green deployment (implantação azul/verde)

Uma estratégia de implantação em que você cria dois ambientes separados, mas idênticos. Você executa a versão atual da aplicação em um ambiente (azul) e a nova versão da aplicação no outro ambiente (verde). Essa estratégia ajuda você a reverter rapidamente com o mínimo de impacto.

bot

Uma aplicação de software que executa tarefas automatizadas na internet e simula a atividade ou interação humana. Alguns bots são úteis ou benéficos, como crawlers da web que indexam informações na internet. Outros bots, conhecidos como bots maliciosos, têm como objetivo causar disrupção ou danos a indivíduos ou organizações.

botnet

Redes de bots infectadas por malware e sob o controle de uma única parte, conhecidas como bot herder ou operador de bots. Os botnets são o mecanismo mais conhecido para escalar bots e seu impacto.

ramo

Uma área contida de um repositório de código. A primeira ramificação criada em um repositório é a ramificação principal. Você pode criar uma nova ramificação a partir de uma ramificação existente e, em seguida, desenvolver recursos ou corrigir bugs na nova ramificação. Uma ramificação que você cria para gerar um recurso é comumente chamada de ramificação de recurso. Quando o recurso estiver pronto para lançamento, você mesclará a ramificação do recurso de volta com a ramificação principal. Para obter mais informações, consulte Sobre filiais (GitHub documentação).

Acesso de emergência

Em circunstâncias excepcionais e por meio de um processo aprovado, um meio rápido para um usuário obter acesso a um Conta da AWS que ele normalmente não tem permissão para acessar. Para obter mais informações, consulte o indicador Implement break-glass procedures nas orientações do AWS Well-Architected.

estratégia brownfield

A infraestrutura existente em seu ambiente. Ao adotar uma estratégia brownfield para uma arquitetura de sistema, você desenvolve a arquitetura de acordo com as restrições dos sistemas e da infraestrutura atuais. Se estiver expandindo a infraestrutura existente, poderá combinar as estratégias brownfield e greenfield.

cache do buffer

A área da memória em que os dados acessados com mais frequência são armazenados.

capacidade de negócios

O que uma empresa faz para gerar valor (por exemplo, vendas, atendimento ao cliente ou marketing). As arquiteturas de microsserviços e as decisões de desenvolvimento podem

ser orientadas por recursos de negócios. Para obter mais informações, consulte a seção Organizados de acordo com as capacidades de negócios do whitepaper Executar microsserviços conteinerizados na AWS.

planejamento de continuidade de negócios (BCP)

Um plano que aborda o impacto potencial de um evento disruptivo, como uma migração em grande escala, nas operações e permite que uma empresa retome as operações rapidamente.

C

CAF

Veja AWS Cloud Adoption Framework.

implantação canário

O lançamento lento e incremental de uma versão para usuários finais. Quando estiver confiante, você implanta a nova versão e substitui a versão atual por completo.

CCoE

Veja Centro de Excelência da Nuvem.

CDC

Veja captura de dados de alteração.

captura de dados de alterações (CDC)

O processo de rastrear alterações em uma fonte de dados, como uma tabela de banco de dados, e registrar metadados sobre a alteração. É possível usar o CDC para várias finalidades, como auditar ou replicar alterações em um sistema de destino para manter a sincronização.

engenharia do caos

Introduzir intencionalmente falhas ou eventos disruptivos para testar a resiliência de um sistema. Você pode usar AWS Fault Injection Service (AWS FIS) para realizar experimentos que estressam suas AWS cargas de trabalho e avaliar sua resposta.

CI/CD

Veja integração e entrega contínuas.

classificação

Um processo de categorização que ajuda a gerar previsões. Os modelos de ML para problemas de classificação predizem um valor discreto. Os valores discretos são sempre diferentes uns dos outros. Por exemplo, um modelo pode precisar avaliar se há ou não um carro em uma imagem.

criptografia no lado do cliente

Criptografia de dados localmente, antes que o alvo os AWS service (Serviço da AWS) receba.

Centro de excelência em nuvem (CCoE)

Uma equipe multidisciplinar que impulsiona os esforços de adoção da nuvem em toda a organização, incluindo o desenvolvimento de práticas recomendas de nuvem, a mobilização de recursos, o estabelecimento de cronogramas de migração e a liderança da organização em transformações em grande escala. Para obter mais informações, consulte as publicações CCo E no blog de estratégia Nuvem AWS corporativa.

computação em nuvem

A tecnologia de nuvem normalmente usada para armazenamento de dados remoto e gerenciamento de dispositivos de IoT. A computação em nuvem é normalmente conectada à tecnologia de computação de borda.

modelo operacional em nuvem

Em uma organização de TI, o modelo operacional usado para criar, amadurecer e otimizar um ou mais ambientes de nuvem. Para obter mais informações, consulte Criar seu modelo operacional de nuvem.

estágios de adoção da nuvem

As quatro fases pelas quais as organizações normalmente passam ao migrar para a Nuvem AWS:

- Projeto: executar alguns projetos relacionados à nuvem para fins de prova de conceito e aprendizado
- Fundação — Fazer investimentos fundamentais para escalar sua adoção da nuvem (por exemplo, criar uma landing zone, definir um CCo E, estabelecer um modelo de operações)
- Migração: migrar aplicações individuais
- Reinvenção: otimizar produtos e serviços e inovar na nuvem

Esses estágios foram definidos por Stephen Orban na postagem do blog The Journey Toward Cloud-First & the Stages of Adoption no blog de estratégia Nuvem AWS empresarial. Para obter

informações sobre como eles se relacionam com a estratégia de AWS migração, consulte o guia de preparação para migração.

CMDB

Veja banco de dados de gerenciamento de configuração.

repositório de código

Um local onde o código-fonte e outros ativos, como documentação, amostras e scripts, são armazenados e atualizados por meio de processos de controle de versão. Os repositórios de nuvem comuns incluem o GitHub ou o Bitbucket Cloud. Cada versão do código é chamada de ramificação. Em uma estrutura de microsserviços, cada repositório é dedicado a uma única peça de funcionalidade. Um único pipeline de CI/CD pode usar vários repositórios.

cache frio

Um cache de buffer que está vazio, não está bem preenchido ou contém dados obsoletos ou irrelevantes. Isso afeta a performance porque a instância do banco de dados deve ler da memória principal ou do disco, um processo que é mais lento do que a leitura do cache do buffer.

dados frios

Dados que raramente são acessados e geralmente são históricos. Ao consultar esse tipo de dados, consultas lentas geralmente são aceitáveis. Mover esses dados para níveis ou classes de armazenamento de baixo desempenho e menos caros pode reduzir os custos.

visão computacional (CV)

Um campo de IA que usa machine learning para analisar e extrair informações de formatos visuais, como vídeos e imagens digitais. Por exemplo, a Amazon SageMaker AI fornece algoritmos de processamento de imagem para CV.

desvio de configuração

Em uma workload, uma alteração de configuração em relação ao estado esperado. Isso pode fazer com que a workload se torne incompatível e, normalmente, é gradual e não intencional.

banco de dados de gerenciamento de configuração (CMDB)

Um repositório que armazena e gerencia informações sobre um banco de dados e seu ambiente de TI, incluindo componentes de hardware e software e suas configurações. Normalmente, os dados de um CMDB são usados no estágio de descoberta e análise do portfólio da migração.

pacote de conformidade

Um conjunto de AWS Config regras e ações de remediação que você pode montar para personalizar suas verificações de conformidade e segurança. Você pode implantar um pacote de conformidade como uma entidade única em uma Conta da AWS região ou em uma organização usando um modelo YAML. Para obter mais informações, consulte Pacotes de conformidade na documentação. AWS Config

integração contínua e entrega contínua (CI/CD)

O processo de automatizar os estágios de origem, criação, teste, preparação e produção do processo de lançamento do software. CI/CD é comumente descrito como um pipeline. CI/CD pode ajudá-lo a automatizar processos, melhorar a produtividade, melhorar a qualidade do código e entregar com mais rapidez. Para obter mais informações, consulte Benefícios da entrega contínua. CD também pode significar implantação contínua. Para obter mais informações, consulte Entrega contínua versus implantação contínua.

CV

Veja visão computacional.

D

dados em repouso

Dados estacionários em sua rede, por exemplo, dados que estão em um armazenamento.

classificação de dados

Um processo para identificar e categorizar os dados em sua rede com base em criticalidade e confidencialidade. É um componente crítico de qualquer estratégia de gerenciamento de riscos de segurança cibernética, pois ajuda a determinar os controles adequados de proteção e retenção para os dados. A classificação de dados é um componente do pilar de segurança no AWS Well-Architected Framework. Para obter mais informações, consulte Classificação de dados.

desvio de dados

Uma variação significativa entre os dados de produção e os dados usados para treinar um modelo de ML ou uma alteração significativa nos dados de entrada ao longo do tempo. O desvio de dados pode reduzir a qualidade geral, a precisão e a imparcialidade das previsões do modelo de ML.

dados em trânsito

Dados que estão se movendo ativamente pela sua rede, como entre os recursos da rede.

data mesh

Um framework de arquitetura que fornece propriedade de dados distribuída e descentralizada com gerenciamento e governança centralizados.

minimização de dados

O princípio de coletar e processar apenas os dados estritamente necessários. Praticar a minimização de dados no Nuvem AWS pode reduzir os riscos de privacidade, os custos e a pegada de carbono de sua análise.

perímetro de dados

Um conjunto de proteções preventivas em seu AWS ambiente que ajudam a garantir que somente identidades confiáveis acessem recursos confiáveis das redes esperadas. Para obter mais informações, consulte Construindo um perímetro de dados em. AWS

pré-processamento de dados

A transformação de dados brutos em um formato que seja facilmente analisado por seu modelo de ML. O pré-processamento de dados pode significar a remoção de determinadas colunas ou linhas e o tratamento de valores ausentes, inconsistentes ou duplicados.

proveniência dos dados

O processo de rastrear a origem e o histórico dos dados ao longo de seu ciclo de vida, por exemplo, como os dados foram gerados, transmitidos e armazenados.

titular dos dados

Um indivíduo cujos dados estão sendo coletados e processados.

data warehouse

Um sistema de gerenciamento de dados compatível com business intelligence, como analytics. Os data warehouses geralmente contêm grandes quantidades de dados históricos e geralmente são usados para consultas e análises.

linguagem de definição de dados (DDL)

Instruções ou comandos para criar ou modificar a estrutura de tabelas e objetos em um banco de dados.

linguagem de manipulação de dados (DML)

Instruções ou comandos para modificar (inserir, atualizar e excluir) informações em um banco de dados.

DDL

Veja linguagem de definição de banco de dados.

deep ensemble

A combinação de vários modelos de aprendizado profundo para gerar previsões. Os deep ensembles podem ser usados para produzir uma previsão mais precisa ou para estimar a incerteza nas previsões.

Aprendizado profundo

Um subcampo do ML que usa várias camadas de redes neurais artificiais para identificar o mapeamento entre os dados de entrada e as variáveis-alvo de interesse.

defense-in-depth

Uma abordagem de segurança da informação na qual uma série de mecanismos e controles de segurança são cuidadosamente distribuídos por toda a rede de computadores para proteger a confidencialidade, a integridade e a disponibilidade da rede e dos dados nela contidos. Ao adotar essa estratégia AWS, você adiciona vários controles em diferentes camadas da AWS Organizations estrutura para ajudar a proteger os recursos. Por exemplo, uma defense-in-depth abordagem pode combinar autenticação multifatorial, segmentação de rede e criptografia.

administrador delegado

Em AWS Organizations, um serviço compatível pode registrar uma conta de AWS membro para administrar as contas da organização e gerenciar as permissões desse serviço. Essa conta é chamada de administrador delegado para esse serviço Para obter mais informações e uma lista de serviços compatíveis, consulte Serviços que funcionam com o AWS Organizations na documentação do AWS Organizations .

implantação

O processo de criar uma aplicação, novos recursos ou correções de código disponíveis no ambiente de destino. A implantação envolve a implementação de mudanças em uma base de código e, em seguida, a criação e execução dessa base de código nos ambientes da aplicação

ambiente de desenvolvimento

Veja ambiente.

controle detectivo

Um controle de segurança projetado para detectar, registrar e alertar após a ocorrência de um evento. Esses controles são uma segunda linha de defesa, alertando você sobre eventos de segurança que contornaram os controles preventivos em vigor. Para obter mais informações, consulte Controles detectivos em Como implementar controles de segurança na AWS.

mapeamento do fluxo de valor de desenvolvimento (DVSM)

Um processo usado para identificar e priorizar restrições que afetam negativamente a velocidade e a qualidade em um ciclo de vida de desenvolvimento de software. O DVSM estende o processo de mapeamento do fluxo de valor originalmente projetado para práticas de manufatura enxuta. Ele se concentra nas etapas e equipes necessárias para criar e movimentar valor por meio do processo de desenvolvimento de software.

gêmeo digital

Uma representação virtual de um sistema real, como um prédio, fábrica, equipamento industrial ou linha de produção. Os gêmeos digitais oferecem suporte à manutenção preditiva, ao monitoramento remoto e à otimização da produção.

tabela de dimensões

Em um esquema em estrela, uma tabela menor que contém atributos de dados sobre dados quantitativos em uma tabela de fatos. Os atributos da tabela de dimensões geralmente são campos de texto ou números discretos que se comportam como texto. Esses atributos normalmente são usados para restringir consultas, filtrar e rotular conjuntos de resultados.

desastre

Um evento que impede que uma workload ou sistema cumpra seus objetivos de negócios em seu local principal de implantação. Esses eventos podem ser desastres naturais, falhas técnicas ou o resultado de ações humanas, como configuração incorreta não intencional ou ataque de malware.

Recuperação de desastres (RD)

A estratégia e o processo que você usa para minimizar o tempo de inatividade e a perda de dados causados por um disastre. Para obter mais informações, consulte Recuperação de desastres de cargas de trabalho em AWS: Recuperação na nuvem no AWS Well-Architected Framework.

DML

Veja linguagem de manipulação de banco de dados.

design orientado por domínio

Uma abordagem ao desenvolvimento de um sistema de software complexo conectando seus componentes aos domínios em evolução, ou principais metas de negócios, atendidos por cada componente. Esse conceito foi introduzido por Eric Evans em seu livro, Design orientado por domínio: lidando com a complexidade no coração do software (Boston: Addison-Wesley Professional, 2003). Para obter informações sobre como usar o design orientado por domínio com o padrão strangler fig, consulte Modernizar incrementalmente os serviços web herdados do Microsoft ASP.NET (ASMX) usando contêineres e o Amazon API Gateway.

DR

Veja recuperação de desastres.

Detecção da oscilação

Rastreamento de desvios de uma configuração de linha de base. Por exemplo, você pode usar AWS CloudFormation para detectar desvios nos recursos do sistema ou AWS Control Tower para detectar mudanças em seu landing zone que possam afetar a conformidade com os requisitos de governança.

DVSM

Veja mapeamento do fluxo de valor de desenvolvimento.

E

EDA

Veja análise exploratória de dados.

EDI

Veja intercâmbio eletrônico de dados.

computação de borda

A tecnologia que aumenta o poder computacional de dispositivos inteligentes nas bordas de uma rede de IoT. Quando comparada com a computação em nuvem, a computação de borda pode reduzir a latência da comunicação e melhorar o tempo de resposta.

intercâmbio eletrônico de dados (EDI)

A troca automatizada de documentos comerciais entre organizações. Para obter mais informações, consulte O que é EDI (Intercâmbio eletrônico de dados)?.

criptografia

Um processo de computação que transforma dados de texto simples, legíveis por humanos, em texto cifrado.

chave de criptografia

Uma sequência criptográfica de bits aleatórios que é gerada por um algoritmo de criptografia. As chaves podem variar em tamanho, e cada chave foi projetada para ser imprevisível e exclusiva.

endianismo

A ordem na qual os bytes são armazenados na memória do computador. Os sistemas big-endian armazenam o byte mais significativo antes. Os sistemas little-endian armazenam o byte menos significativo antes.

endpoint

Veja endpoint de serviço.

serviço de endpoint

Um serviço que pode ser hospedado em uma nuvem privada virtual (VPC) para ser compartilhado com outros usuários. Você pode criar um serviço de endpoint com AWS PrivateLink e conceder permissões a outros diretores Contas da AWS ou a AWS Identity and Access Management (IAM). Essas contas ou entidades principais podem se conectar ao serviço de endpoint de maneira privada criando endpoints da VPC de interface. Para obter mais informações, consulte Criar um serviço de endpoint na documentação do Amazon Virtual Private Cloud (Amazon VPC).

planejamento de recursos empresariais (ERP)

Um sistema que automatiza e gerencia os principais processos de negócios (como contabilidade, MES e gerenciamento de projetos) para uma empresa.

criptografia envelopada

O processo de criptografar uma chave de criptografia com outra chave de criptografia. Para obter mais informações, consulte Criptografia de envelope na documentação AWS Key Management Service (AWS KMS).

ambiente

Uma instância de uma aplicação em execução. Estes são tipos comuns de ambientes na computação em nuvem:

- ambiente de desenvolvimento: uma instância de uma aplicação em execução que está disponível somente para a equipe principal responsável pela manutenção da aplicação. Ambientes de desenvolvimento são usados para testar mudanças antes de promovê-las para ambientes superiores. Esse tipo de ambiente às vezes é chamado de ambiente de teste.
- ambientes inferiores: todos os ambientes de desenvolvimento para uma aplicação, como aqueles usados para compilações e testes iniciais.
- ambiente de produção: uma instância de uma aplicação em execução que os usuários finais podem acessar. Em um CI/CD pipeline, o ambiente de produção é o último ambiente de implantação.
- ambientes superiores: todos os ambientes que podem ser acessados por usuários que não sejam a equipe principal de desenvolvimento. Isso pode incluir um ambiente de produção, ambientes de pré-produção e ambientes para testes de aceitação do usuário.

epic

Em metodologias ágeis, categorias funcionais que ajudam a organizar e priorizar seu trabalho. Os epics fornecem uma descrição de alto nível dos requisitos e das tarefas de implementação. Por exemplo, os épicos de segurança AWS da CAF incluem gerenciamento de identidade e acesso, controles de detetive, segurança de infraestrutura, proteção de dados e resposta a incidentes. Para obter mais informações sobre epics na estratégia de migração da AWS , consulte o guia de implementação do programa.

ERP

Veja planejamento de recursos empresariais.

análise exploratória de dados (EDA)

O processo de analisar um conjunto de dados para entender suas principais características. Você coleta ou agrega dados e, em seguida, realiza investigações iniciais para encontrar padrões, detectar anomalias e verificar suposições. O EDA é realizado por meio do cálculo de estatísticas resumidas e da criação de visualizações de dados.

F

tabela de fatos

A tabela central em um esquema em estrela. Ela armazena dados quantitativos sobre as operações comerciais. Normalmente, uma tabela de fatos contém dois tipos de colunas: as que contêm medidas e as que contêm uma chave externa para uma tabela de dimensões.

Antecipar-se à falha

Uma filosofia que usa testes frequentes e incrementais para reduzir o ciclo de vida do desenvolvimento. É uma parte essencial de uma abordagem ágil.

delimitação de isolamento contra falhas

No Nuvem AWS, um limite, como uma zona de disponibilidade, Região da AWS um plano de controle ou um plano de dados, que limita o efeito de uma falha e ajuda a melhorar a resiliência das cargas de trabalho. Para obter mais informações, consulte AWS Fault Isolation Boundaries.

ramificação de recursos

Veja ramificação.

recursos

Os dados de entrada usados para fazer uma previsão. Por exemplo, em um contexto de manufatura, os recursos podem ser imagens capturadas periodicamente na linha de fabricação.

importância do recurso

O quanto um recurso é importante para as previsões de um modelo. Isso geralmente é expresso como uma pontuação numérica que pode ser calculada por meio de várias técnicas, como Shapley Additive Explanations (SHAP) e gradientes integrados. Para obter mais informações, consulte Interpretabilidade do modelo de aprendizado de máquina com AWS.

transformação de recursos

O processo de otimizar dados para o processo de ML, incluindo enriquecer dados com fontes adicionais, escalar valores ou extrair vários conjuntos de informações de um único campo de dados. Isso permite que o modelo de ML se beneficie dos dados. Por exemplo, se a data “2021-05-27 00:15:37” for dividida em “2021”, “maio”, “quinta” e “15”, isso poderá ajudar o algoritmo de aprendizado a aprender padrões diferenciados associados a diferentes componentes de dados.

prompt few shot

Fornecer a um LLM um pequeno número de exemplos que demonstram a tarefa e o resultado desejado antes de solicitar que ele execute uma tarefa semelhante. Essa técnica é uma aplicação do aprendizado em contexto, em que os modelos aprendem com exemplos (shots) incorporados aos prompts. Prompts few-shot podem ser eficazes para tarefas que exigem formatação, raciocínio ou conhecimento de domínio específicos. Veja também prompts zero-shot.

FGAC

Veja controle de acesso refinado.

Controle de acesso refinado (FGAC)

O uso de várias condições para permitir ou negar uma solicitação de acesso.

migração flash-cut

Um método de migração de banco de dados que usa replicação contínua de dados via captura de dados de alteração para migrar os dados no menor tempo possível, em vez de usar uma abordagem em fases. O objetivo é reduzir ao mínimo o tempo de inatividade.

FM

Veja modelo de base.

modelo de base (FM)

Uma grande rede neural de aprendizado profundo que vem treinando em grandes conjuntos de dados generalizados e não rotulados. FMs são capazes de realizar uma ampla variedade de tarefas gerais, como entender a linguagem, gerar texto e imagens e conversar em linguagem natural. Para obter mais informações, consulte O que são modelos de base?.

G

IA generativa

Um subconjunto de modelos de IA que foram treinados em grandes quantidades de dados e que podem usar um simples prompt de texto para criar novos artefatos e conteúdo, como imagens, vídeos, texto e áudio. Para obter mais informações, consulte O que é IA generativa?.

bloqueio geográfico

Veja restrições geográficas.

restrições geográficas (bloqueio geográfico)

Na Amazon CloudFront, uma opção para impedir que usuários em países específicos acessem distribuições de conteúdo. É possível usar uma lista de permissões ou uma lista de bloqueios para especificar países aprovados e banidos. Para obter mais informações, consulte Restringir a distribuição geográfica do seu conteúdo na CloudFront documentação.

Fluxo de trabalho do GitFlow

Uma abordagem na qual ambientes inferiores e superiores usam ramificações diferentes em um repositório de código-fonte. O fluxo de trabalho do Gitflow é considerado legado, e o fluxo de trabalho trunk-based é a abordagem moderna e preferencial.

golden image

Um snapshot de um sistema ou software usado como modelo para implantar novas instâncias desse sistema ou software. Por exemplo, na manufatura, uma golden image pode ser usada para provisionar software em vários dispositivos e ajudar a melhorar a velocidade, a escalabilidade e a produtividade nas operações de fabricação de dispositivos.

estratégia greenfield

A ausência de infraestrutura existente em um novo ambiente. Ao adotar uma estratégia greenfield para uma arquitetura de sistema, é possível selecionar todas as novas tecnologias sem a restrição da compatibilidade com a infraestrutura existente, também conhecida como brownfield. Se estiver expandindo a infraestrutura existente, poderá combinar as estratégias brownfield e greenfield.

barreira de proteção

Uma regra de alto nível que ajuda a governar recursos, políticas e conformidade em todas as unidades organizacionais ()OUs. Barreiras de proteção preventivas impõem políticas para garantir o alinhamento a padrões de conformidade. Elas são implementadas usando políticas de controle de serviço e limites de permissões do IAM. Barreiras de proteção detectivas detectam violações de políticas e problemas de conformidade e geram alertas para remediação. Eles são implementados usando AWS Config, AWS Security Hub CSPM, Amazon GuardDuty AWS Trusted Advisor, Amazon Inspector e verificações personalizadas AWS Lambda .

H

HA

Veja alta disponibilidade.

migração heterogênea de bancos de dados

Migrar seu banco de dados de origem para um banco de dados de destino que usa um mecanismo de banco de dados diferente (por exemplo, Oracle para Amazon Aurora). A migração heterogênea geralmente faz parte de um esforço de redefinição da arquitetura, e converter

o esquema pode ser uma tarefa complexa. O AWS fornece o AWS SCT para ajudar nas conversões de esquemas.

alta disponibilidade (HA)

A capacidade de uma workload operar continuamente, sem intervenção, em caso de desafios ou desastres. Os sistemas AH são projetados para realizar o failover automático, oferecer consistentemente desempenho de alta qualidade e lidar com diferentes cargas e falhas com impacto mínimo no desempenho.

modernização de historiador

Uma abordagem usada para modernizar e atualizar os sistemas de tecnologia operacional (OT) para melhor atender às necessidades do setor de manufatura. Um historiador é um tipo de banco de dados usado para coletar e armazenar dados de várias fontes em uma fábrica.

dados de hold-out

Uma parte dos dados históricos rotulados que são retidos de um conjunto de dados usado para treinar um modelo de machine learning. Você pode usar dados de hold-out para avaliar a performance do modelo comparando as predições do modelo com os dados de retenção.

migração homogênea de bancos de dados

Migrar seu banco de dados de origem para um banco de dados de destino que compartilha o mesmo mecanismo de banco de dados (por exemplo, Microsoft SQL Server para Amazon RDS para SQL Server). A migração homogênea geralmente faz parte de um esforço de redefinição da hospedagem ou da plataforma. É possível usar utilitários de banco de dados nativos para migrar o esquema.

dados quentes

Dados acessados com frequência, como dados em tempo real ou dados translacionais recentes. Esses dados normalmente exigem uma camada ou classe de armazenamento de alto desempenho para fornecer respostas rápidas às consultas.

hotfix

Uma correção urgente para um problema crítico em um ambiente de produção. Devido à sua urgência, um hotfix geralmente é feito fora do fluxo de trabalho normal de DevOps lançamento.

período de hipercuidados

Imediatamente após a substituição, o período em que uma equipe de migração gerencia e monitora as aplicações migradas na nuvem para resolver quaisquer problemas. Normalmente,

a duração desse período é de 1 a 4 dias. No final do período de hipercuidados, a equipe de migração normalmente transfere a responsabilidade pelas aplicações para a equipe de operações de nuvem.

eu

IaC

Veja infraestrutura como código.

Política baseada em identidade

Uma política anexada a um ou mais diretores do IAM que define suas permissões no Nuvem AWS ambiente.

aplicação ociosa

Uma aplicação que tem um uso médio de CPU e memória entre 5 e 20% em um período de 90 dias. Em um projeto de migração, é comum retirar essas aplicações ou retê-las on-premises.

IIoT

Veja Internet das Coisas Industrial.

infraestrutura imutável

Um modelo que implanta uma nova infraestrutura para workloads de produção em vez de atualizar, aplicar patches ou modificar a infraestrutura existente. Infraestruturas imutáveis são inerentemente mais consistentes, confiáveis e preditivas do que infraestruturas mutáveis. Para obter mais informações, consulte a prática recomendada Implantar usando infraestrutura imutável no AWS Well-Architected Framework.

VPC de entrada (admissão)

Em uma arquitetura de AWS várias contas, uma VPC que aceita, inspeciona e roteia conexões de rede de fora de um aplicativo. A Arquitetura de Referência de AWS Segurança recomenda configurar sua conta de rede com entrada, saída e inspeção VPCs para proteger a interface bidirecional entre seu aplicativo e a Internet em geral.

migração incremental

Uma estratégia de substituição na qual você migra a aplicação em pequenas partes, em vez de realizar uma única substituição completa. Por exemplo, é possível mover inicialmente

apenas alguns microsserviços ou usuários para o novo sistema. Depois de verificar se tudo está funcionando corretamente, mova os microsserviços ou usuários adicionais de forma incremental até poder descomissionar seu sistema herdado. Essa estratégia reduz os riscos associados a migrações de grande porte.

Indústria 4.0

Um termo que foi introduzido por Klaus Schwab em 2016 para se referir à modernização dos processos de manufatura por meio de avanços em conectividade, dados em tempo real, automação, analytics e IA/ML.

infraestrutura

Todos os recursos e ativos contidos no ambiente de uma aplicação.

Infraestrutura como código (IaC)

O processo de provisionamento e gerenciamento da infraestrutura de uma aplicação por meio de um conjunto de arquivos de configuração. A IaC foi projetada para ajudar você a centralizar o gerenciamento da infraestrutura, padronizar recursos e escalar rapidamente para que novos ambientes sejam reproduzíveis, confiáveis e consistentes.

Internet industrial das coisas (IIoT)

O uso de sensores e dispositivos conectados à Internet nos setores industriais, como manufatura, energia, automotivo, saúde, ciências biológicas e agricultura. Para obter mais informações, consulte Criando uma estratégia de transformação digital industrial da Internet das Coisas (IIoT).

VPC de inspeção

Em uma arquitetura de AWS várias contas, uma VPC centralizada que gerencia as inspeções do tráfego de rede entre VPCs (na mesma ou em diferentes Regiões da AWS) a Internet e as redes locais. A Arquitetura de Referência de AWS Segurança recomenda configurar sua conta de rede com entrada, saída e inspeção VPCs para proteger a interface bidirecional entre seu aplicativo e a Internet em geral.

Internet das coisas (IoT)

A rede de objetos físicos conectados com sensores ou processadores incorporados que se comunicam com outros dispositivos e sistemas pela Internet ou por uma rede de comunicação local. Para obter mais informações, consulte O que é IoT?

interpretabilidade

Uma característica de um modelo de machine learning que descreve o grau em que um ser humano pode entender como as previsões do modelo dependem de suas entradas. Para obter mais informações, consulte Interpretabilidade do modelo de aprendizado de máquina com AWS.

IoT

Veja Internet das Coisas.

Biblioteca de informações de TI (ITIL)

Um conjunto de práticas recomendadas para fornecer serviços de TI e alinhar esses serviços a requisitos de negócios. A ITIL fornece a base para o ITSM.

Gerenciamento de serviços de TI (ITSM)

Atividades associadas a design, implementação, gerenciamento e suporte de serviços de TI para uma organização. Para obter informações sobre a integração de operações em nuvem com ferramentas de ITSM, consulte o guia de integração de operações.

ITIL

Veja biblioteca de informações de TI.

ITSM

Veja gerenciamento de serviços de TI.

L

controle de acesso baseado em etiqueta (LBAC)

Uma implementação do controle de acesso obrigatório (MAC) em que os usuários e os dados em si recebem explicitamente um valor de etiqueta de segurança. A interseção entre a etiqueta de segurança do usuário e a etiqueta de segurança dos dados determina quais linhas e colunas podem ser vistas pelo usuário.

zona de pouso

Uma landing zone é um AWS ambiente bem arquitetado, com várias contas, escalável e seguro. Um ponto a partir do qual suas organizações podem iniciar e implantar rapidamente workloads e aplicações com confiança em seu ambiente de segurança e infraestrutura. Para obter mais

informações sobre zonas de pouso, consulte Configurar um ambiente da AWS com várias contas seguro e escalável.

grande modelo de linguagem (LLM)

Um modelo de IA de aprendizado profundo pré-treinado em uma grande quantidade de dados. Um LLM pode realizar várias tarefas, como responder a perguntas, resumir documentos, traduzir texto para outros idiomas e completar frases. Para obter mais informações, consulte O que são LLMs.

migração de grande porte

Uma migração de 300 servidores ou mais.

LBAC

Veja controle de acesso baseado em rótulo.

privilégio mínimo

A prática recomendada de segurança de conceder as permissões mínimas necessárias para executar uma tarefa. Para obter mais informações, consulte Aplicar permissões de privilégios mínimos na documentação do IAM.

mover sem alterações (lift-and-shift)

Veja 7 Rs.

sistema little-endian

Um sistema que armazena o byte menos significativo antes. Veja também endianness.

LLM

Veja grande modelo de linguagem.

ambientes inferiores

Veja ambiente.

M

machine learning (ML)

Um tipo de inteligência artificial que usa algoritmos e técnicas para reconhecimento e aprendizado de padrões. O ML analisa e aprende com dados gravados, por exemplo, dados da

Internet das Coisas (IoT), para gerar um modelo estatístico baseado em padrões. Para obter mais informações, consulte Machine learning.

ramificação principal

Veja ramificação.

Malware

Software projetado para comprometer a segurança ou a privacidade do computador. O malware pode interromper os sistemas do computador, vazar informações sensíveis ou obter acesso não autorizado. Exemplos de malware incluem vírus, worms, ransomware, cavalos de Troia, spyware e keyloggers.

Serviços gerenciados

Serviços da AWS para o qual AWS opera a camada de infraestrutura, o sistema operacional e as plataformas, e você acessa os endpoints para armazenar e recuperar dados. O Amazon Simple Storage Service (Amazon S3) e o Amazon DynamoDB são exemplos de serviços gerenciados. Eles também são conhecidos como serviços abstraídos.

sistema de execução de manufatura (MES)

Um sistema de software para rastrear, monitorar, documentar e controlar processos de produção que convertem matérias-primas em produtos acabados no chão de fábrica.

MAP

Veja Programa de Aceleração da Migração.

mecanismo

Um processo completo em que você cria uma ferramenta, impulsiona a adoção da ferramenta e, em seguida, inspeciona os resultados para fazer ajustes. Um mecanismo é um ciclo que se reforça e se aprimora à medida que opera. Para obter mais informações, consulte Construindo mecanismos no AWS Well-Architected Framework.

conta de membro

Todos, Contas da AWS exceto a conta de gerenciamento, que fazem parte de uma organização em AWS Organizations. Uma conta só pode ser membro de uma organização de cada vez.

MES

Veja sistema de execução de manufatura.

Transporte de Telemetria de Enfileiramento de Mensagens (MQTT)

Um protocolo de comunicação leve machine-to-machine (M2M), baseado no padrão de publicação/assinatura, para dispositivos de IoT com recursos limitados.

microsserviço

Um serviço pequeno e independente que se comunica de forma bem definida APIs e normalmente é de propriedade de equipes pequenas e independentes. Por exemplo, um sistema de seguradora pode incluir microsserviços que mapeiam as capacidades comerciais, como vendas ou marketing, ou subdomínios, como compras, reclamações ou análises. Os benefícios dos microsserviços incluem agilidade, escalabilidade flexível, fácil implantação, código reutilizável e resiliência. Para obter mais informações, consulte Integração de microsserviços usando serviços sem AWS servidor.

arquitetura de microsserviços

Uma abordagem à criação de aplicações com componentes independentes que executam cada processo de aplicação como um microsserviço. Esses microsserviços se comunicam por meio de uma interface bem definida usando leveza. APIs Cada microsserviço nessa arquitetura pode ser atualizado, implantado e escalado para atender à demanda por funções específicas de uma aplicação. Para obter mais informações, consulte Implementação de microsserviços em. AWS

Programa de Aceleração da Migração (MAP)

Um AWS programa que fornece suporte de consultoria, treinamento e serviços para ajudar as organizações a criar uma base operacional sólida para migrar para a nuvem e ajudar a compensar o custo inicial das migrações. O MAP inclui uma metodologia de migração para executar migrações legadas de forma metódica e um conjunto de ferramentas para automatizar e acelerar cenários comuns de migração.

migração em escala

O processo de mover a maior parte do portfólio de aplicações para a nuvem em ondas, com mais aplicações sendo movidas em um ritmo mais rápido a cada onda. Essa fase usa as práticas recomendadas e lições aprendidas nas fases anteriores para implementar uma fábrica de migração de equipes, ferramentas e processos para agilizar a migração de workloads por meio de automação e entrega ágeis. Esta é a terceira fase da estratégia de migração para a AWS.

fábrica de migração

Equipes multifuncionais que simplificam a migração de workloads por meio de abordagens automatizadas e ágeis. As equipes da fábrica de migração geralmente incluem operações,

analistas e proprietários de negócios, engenheiros de migração, desenvolvedores e DevOps profissionais que trabalham em sprints. Entre 20 e 50% de um portfólio de aplicações corporativas consiste em padrões repetidos que podem ser otimizados por meio de uma abordagem de fábrica. Para obter mais informações, consulte discussão sobre fábricas de migração e o guia do Cloud Migration Factory neste conjunto de conteúdo.

metadados de migração

As informações sobre a aplicação e o servidor necessárias para concluir a migração. Cada padrão de migração exige um conjunto de metadados de migração diferente. Exemplos de metadados de migração incluem a sub-rede, o grupo de segurança e AWS a conta de destino.

padrão de migração

Uma tarefa de migração repetível que detalha a estratégia de migração, o destino da migração e a aplicação ou o serviço de migração usado. Exemplo: rehospede a migração para o Amazon EC2 AWS com o Application Migration Service.

Avaliação de Portfólio para Migração (MPA)

Uma ferramenta on-line que fornece informações para validar o caso de negócios para migrar para a Nuvem AWS. O MPA fornece avaliação detalhada do portfólio (dimensionamento correto do servidor, preços, comparações de TCO, análise de custos de migração), bem como planejamento de migração (análise e coleta de dados de aplicações, agrupamento de aplicações, priorização de migração e planejamento de ondas). A ferramenta MPA (requer login) está disponível gratuitamente para todos os AWS consultores e consultores parceiros da APN.

Avaliação de Preparação para Migração (MRA)

O processo de obter insights sobre o status de prontidão de uma organização para a nuvem, identificar pontos fortes e fracos e criar um plano de ação para fechar as lacunas identificadas, usando o CAF. AWS Para mais informações, consulte o guia de preparação para migração. A MRA é a primeira fase da estratégia de migração para a AWS.

estratégia de migração

A abordagem usada para migrar uma workload para a Nuvem AWS. Para obter mais informações, veja a entrada 7 Rs neste glossário e consulte Mobilize sua organização para acelerar migrações em grande escala.

ML

Veja machine learning.

modernização

Transformar uma aplicação desatualizada (herdada ou monolítica) e sua infraestrutura em um sistema ágil, elástico e altamente disponível na nuvem para reduzir custos, ganhar eficiência e aproveitar as inovações. Para obter mais informações, consulte Strategy for modernizing applications in the Nuvem AWS.

avaliação de preparação para modernização

Uma avaliação que ajuda a determinar a preparação para modernização das aplicações de uma organização. Ela identifica benefícios, riscos e dependências e determina o quão bem a organização pode acomodar o estado futuro dessas aplicações. O resultado da avaliação é um esquema da arquitetura de destino, um roteiro que detalha as fases de desenvolvimento e os marcos do processo de modernização e um plano de ação para abordar as lacunas identificadas. Para obter mais informações, consulte Evaluating modernization readiness for applications in the Nuvem AWS.

aplicações monolíticas (monólitos)

Aplicações que são executadas como um único serviço com processos fortemente acoplados. As aplicações monolíticas apresentam várias desvantagens. Se um recurso da aplicação apresentar um aumento na demanda, toda a arquitetura deverá ser escalada. Adicionar ou melhorar os recursos de uma aplicação monolítica também se torna mais complexo quando a base de código cresce. Para resolver esses problemas, é possível criar uma arquitetura de microsserviços. Para obter mais informações, consulte Decompor monólitos em microsserviços.

MPA

Veja Avaliação do Portfólio para Migração.

MQTT

Veja Transporte de Telemetria de Enfileiramento de Mensagens.

classificação multiclasse

Um processo que ajuda a gerar previsões para várias classes (prevendo um ou mais de dois resultados). Por exemplo, um modelo de ML pode perguntar “Este produto é um livro, um carro ou um telefone?” ou “Qual categoria de produtos é mais interessante para este cliente?”

infraestrutura mutável

Um modelo que atualiza e modifica a infraestrutura existente para workloads de produção. Para melhorar a consistência, confiabilidade e previsibilidade, o AWS Well-Architected Framework recomenda o uso de infraestrutura imutável como uma prática recomendada.

O

OAC

Veja controle de acesso de origem.

OAI

Veja identidade de acesso de origem.

OCM

Veja gerenciamento de alterações organizacionais.

migração offline

Um método de migração no qual a workload de origem é desativada durante o processo de migração. Esse método envolve tempo de inatividade prolongado e geralmente é usado para workloads pequenas e não críticas.

OI

Veja integração de operações.

Ola

Veja acordo de nível operacional.

migração online

Um método de migração no qual a workload de origem é copiada para o sistema de destino sem ser colocada offline. As aplicações conectadas à workload podem continuar funcionando durante a migração. Esse método envolve um tempo de inatividade nulo ou mínimo e normalmente é usado para workloads essenciais para a produção.

OPC-UA

Veja Open Process Communications - Unified Architecture.

Open Process Communications - Unified Architecture (OPC-UA)

Um protocolo de comunicação machine-to-machine (M2M) para automação industrial. O OPC-UA fornece um padrão de interoperabilidade com esquemas de criptografia, autenticação e autorização de dados.

acordo de nível operacional (OLA)

Um acordo que esclarece o que os grupos funcionais de TI prometem oferecer uns aos outros para apoiar um acordo de serviço (SLA).

análise de prontidão operacional (ORR)

Uma lista de verificação de perguntas e práticas recomendadas associadas que ajudam você a entender, avaliar, prevenir ou reduzir o escopo de incidentes e possíveis falhas. Para obter mais informações, consulte Operational Readiness Reviews (ORR) no AWS Well-Architected Framework.

tecnologia operacional (TO)

Sistemas de hardware e software que trabalham com o ambiente físico para controlar operações, equipamentos e infraestrutura industriais. Na manufatura, a integração dos sistemas de tecnologia da informação (TI) e tecnologia operacional (TO) é o foco principal das transformações da Indústria 4.0.

integração de operações (OI)

O processo de modernização das operações na nuvem, que envolve planejamento de preparação, automação e integração. Para obter mais informações, consulte o guia de integração de operações.

trilha organizacional

Uma trilha criada por ela AWS CloudTrail registra todos os eventos de todos Contas da AWS em uma organização em AWS Organizations. Essa trilha é criada em cada Conta da AWS que faz parte da organização e monitora a atividade em cada conta. Para obter mais informações, consulte Criação de uma trilha para uma organização na CloudTrail documentação.

gerenciamento de alterações organizacionais (OCM)

Uma estrutura para gerenciar grandes transformações de negócios disruptivas de uma perspectiva de pessoas, cultura e liderança. O OCM ajuda as organizações a se prepararem e fazerem a transição para novos sistemas e estratégias, acelerando a adoção de alterações, abordando questões de transição e promovendo mudanças culturais e organizacionais. Na estratégia de AWS migração, essa estrutura é chamada de aceleração de pessoas, devido à velocidade de mudança exigida nos projetos de adoção da nuvem. Para obter mais informações, consulte o guia do OCM.

controle de acesso de origem (OAC)

Em CloudFront, uma opção aprimorada para restringir o acesso para proteger seu conteúdo do Amazon Simple Storage Service (Amazon S3). O OAC oferece suporte a todos os buckets S3 Regiões da AWS, criptografia do lado do servidor com AWS KMS (SSE-KMS) e solicitações dinâmicas ao bucket S3. `PUT` `DELETE`

Identidade do acesso de origem (OAI)

Em CloudFront, uma opção para restringir o acesso para proteger seu conteúdo do Amazon S3. Quando você usa o OAI, CloudFront cria um principal com o qual o Amazon S3 pode se autenticar. Os diretores autenticados podem acessar o conteúdo em um bucket do S3 somente por meio de uma distribuição específica. CloudFront Veja também OAC, que fornece um controle de acesso mais granular e aprimorado.

ORR

Veja análise de prontidão operacional.

OT

Veja tecnologia operacional.

VPC de saída (egresso)

Em uma arquitetura de AWS várias contas, uma VPC que gerencia conexões de rede que são iniciadas de dentro de um aplicativo. A Arquitetura de Referência de AWS Segurança recomenda configurar sua conta de rede com entrada, saída e inspeção VPCs para proteger a interface bidirecional entre seu aplicativo e a Internet em geral.

P

limite de permissões

Uma política de gerenciamento do IAM anexada a entidades principais do IAM para definir as permissões máximas que o usuário ou perfil podem ter. Para obter mais informações, consulte Limites de permissões na documentação do IAM.

Informações de identificação pessoal (PII)

Informações que, quando visualizadas diretamente ou combinadas com outros dados relacionados, podem ser usadas para inferir razoavelmente a identidade de um indivíduo. Exemplos de PII incluem nomes, endereços e informações de contato.

PII

Veja informações de identificação pessoal.

manual

Um conjunto de etapas predefinidas que capturam o trabalho associado às migrações, como a entrega das principais funções operacionais na nuvem. Um manual pode assumir a forma de scripts, runbooks automatizados ou um resumo dos processos ou etapas necessários para operar seu ambiente modernizado.

PLC

Veja controlador lógico programável.

PLM

Veja gerenciamento do ciclo de vida do produto.

política

Um objeto que pode definir permissões (veja política baseada em identidade), especificar condições de acesso (veja políta baseada em recurso) ou definir as permissões máximas para todas as contas em uma organização no AWS Organizations (veja política de controle de serviços).

persistência poliglota

Escolher de forma independente a tecnologia de armazenamento de dados de um microsserviço com base em padrões de acesso a dados e outros requisitos. Se seus microsserviços tiverem a mesma tecnologia de armazenamento de dados, eles poderão enfrentar desafios de implementação ou apresentar baixa performance. Os microsserviços serão implementados com mais facilidade e alcançarão performance e escalabilidade melhores se usarem o armazenamento de dados mais bem adaptado às suas necessidades.

avaliação do portfólio

Um processo de descobrir, analisar e priorizar o portfólio de aplicações para planejar a migração. Para obter mais informações, consulte Avaliar a preparação para a migração.

predicado

Uma condição de consulta que retorna `true` ou `false`, normalmente localizada em uma cláusula `WHERE`.

pushdown de predicados

Uma técnica de otimização de consultas de banco de dados que filtra os dados na consulta antes da transferência. Isso reduz a quantidade de dados que devem ser recuperados e processados do banco de dados relacional e melhora a performance das consultas.

controle preventivo

Um controle de segurança projetado para evitar que um evento ocorra. Esses controles são a primeira linha de defesa para ajudar a evitar acesso não autorizado ou alterações indesejadas em sua rede. Para obter mais informações, consulte Controles preventivos em Como implementar controles de segurança na AWS.

principal (entidade principal)

Uma entidade AWS que pode realizar ações e acessar recursos. Essa entidade geralmente é um usuário raiz para um Conta da AWS, uma função do IAM ou um usuário. Para obter mais informações, consulte Entidade principal em Termos e conceitos de perfis na documentação do IAM.

Privacidade por design

Uma abordagem em engenharia de sistemas que leva em consideração a privacidade em todo o processo de desenvolvimento.

zonas hospedadas privadas

Um contêiner que contém informações sobre como você deseja que o Amazon Route 53 responda às consultas de DNS para um domínio e seus subdomínios em um ou mais. VPCs Para obter mais informações, consulte Como trabalhar com zonas hospedadas privadas na documentação do Route 53.

controle proativo

Um controle de segurança desenvolvido para evitar a implantação de recursos não conformes. Esses controles verificam os recursos antes de serem provisionados. Se o recurso não estiver em conformidade com o controle, ele não será provisionado. Para obter mais informações, consulte o guia de referência de controles na AWS Control Tower documentação e consulte Controles proativos em Implementação de controles de segurança em AWS.

gerenciamento do ciclo de vida do produto (PLM)

O gerenciamento de dados e processos de um produto em todo o seu ciclo de vida, desde a concepção, o desenvolvimento e o lançamento, passando pelo crescimento e maturidade, até o declínio e a remoção.

ambiente de produção

Veja ambiente.

controlador lógico programável (PLC)

Na manufatura, um computador altamente confiável e adaptável que monitora as máquinas e automatiza os processos de fabricação.

encadeamento de prompts

Uso da saída de um prompt do LLM como entrada para o próximo prompt para gerar respostas melhores. Essa técnica é usada para dividir uma tarefa complexa em subtarefas, ou para refinar ou expandir iterativamente uma resposta preliminar. Isso ajuda a melhorar a precisão e a relevância das respostas de um modelo e permite resultados mais granulares e personalizados.

pseudonimização

O processo de substituir identificadores pessoais em um conjunto de dados por valores de espaço reservado. A pseudonimização pode ajudar a proteger a privacidade pessoal. Os dados pseudonimizados ainda são considerados dados pessoais.

publish/subscribe (pub/sub)

Um padrão que permite comunicações assíncronas entre microsserviços para melhorar a escalabilidade e a capacidade de resposta. Por exemplo, em um MES baseado em microsserviços, um microsserviço pode publicar mensagens de eventos em um canal em que outros microsserviços possam assinar. O sistema pode adicionar novos microsserviços sem alterar o serviço de publicação.

Q

plano de consulta

Uma série de etapas, como instruções, usadas para acessar os dados em um sistema de banco de dados relacional SQL.

regressão de planos de consultas

Quando um otimizador de serviço de banco de dados escolhe um plano menos adequado do que escolhia antes de uma determinada alteração no ambiente de banco de dados ocorrer. Isso pode ser causado por alterações em estatísticas, restrições, configurações do ambiente, associações de parâmetros de consulta e atualizações do mecanismo de banco de dados.

R

Matriz RACI

Veja responsável, aprovador, consultado, informado (RACI).

RAG

Veja geração aumentada via recuperação.

ransomware

Um software mal-intencionado desenvolvido para bloquear o acesso a um sistema ou dados de computador até que um pagamento seja feito.

Matriz RASCI

Veja responsável, aprovador, consultado, informado (RACI).

RCAC

Veja controle de acesso por linha e coluna.

réplica de leitura

Uma cópia de um banco de dados usada somente para leitura. É possível encaminhar consultas para a réplica de leitura e reduzir a carga no banco de dados principal.

Redefinir arquitetura

Veja 7 Rs.

objetivo de ponto de recuperação (RPO).

O máximo período de tempo aceitável desde o último ponto de recuperação de dados. Isso determina o que é considerado uma perda aceitável de dados entre o último ponto de recuperação e a interrupção do serviço.

objetivo de tempo de recuperação (RTO)

O máximo atraso aceitável entre a interrupção e a restauração do serviço.

refatorar

Veja 7 Rs.

Região

Uma coleção de AWS recursos em uma área geográfica. Cada um Região da AWS é isolado e independente dos outros para fornecer tolerância a falhas, estabilidade e resiliência. Para obter informações, consulte Specify which Regiões da AWS your account can use.

regressão

Uma técnica de ML que prevê um valor numérico. Por exemplo, para resolver o problema de "Por qual preço esta casa será vendida?" um modelo de ML pode usar um modelo de regressão linear para prever o preço de venda de uma casa com base em fatos conhecidos sobre a casa (por exemplo, a metragem quadrada).

redefinir a hospedagem

Veja 7 Rs.

versão

Em um processo de implantação, o ato de promover mudanças em um ambiente de produção.

realocar

Veja 7 Rs.

redefinir a plataforma

Veja 7 Rs.

recomprar

Veja 7 Rs.

resiliência

A capacidade de uma aplicação de resistir ou se recuperar de interrupções. Alta disponibilidade e recuperação de desastres são considerações comuns ao planejar a resiliência na Nuvem AWS. Para obter mais informações, consulte Nuvem AWS Resilience.

política baseada em recurso

Uma política associada a um recurso, como um bucket do Amazon S3, um endpoint ou uma chave de criptografia. Esse tipo de política especifica quais entidades principais têm acesso permitido, ações válidas e quaisquer outras condições que devem ser atendidas.

matriz responsável, accountable, consultada, informada (RACI)

Uma matriz que define as funções e responsabilidades de todas as partes envolvidas nas atividades de migração e nas operações de nuvem. O nome da matriz é derivado dos tipos de responsabilidade definidos na matriz: responsável (R), responsabilizável (A), consultado (C) e informado (I). O tipo de suporte (S) é opcional. Se você incluir suporte, a matriz será chamada de matriz RASCI e, se excluir, será chamada de matriz RACI.

controle responsivo

Um controle de segurança desenvolvido para conduzir a remediação de eventos adversos ou desvios em relação à linha de base de segurança. Para obter mais informações, consulte Controles responsivos em Como implementar controles de segurança na AWS.

reter

Veja 7 Rs.

Retirada

Veja 7 Rs.

Geração Aumentada de Recuperação (RAG)

Uma tecnologia de IA generativa em que um LLM faz referência a uma fonte de dados autorizada que está fora de suas fontes de dados de treinamento antes de gerar uma resposta. Por exemplo, um modelo RAG pode realizar uma pesquisa semântica na base de conhecimento ou nos dados personalizados de uma organização. Para obter mais informações, consulte O que é RAG (geração aumentada via recuperação)?.

alternância

O processo de atualizar periodicamente um segredo para dificultar o acesso de um invasor às credenciais.

controle de acesso por linha e coluna (RCAC)

O uso de expressões SQL básicas e flexíveis que tenham regras de acesso definidas. O RCAC consiste em permissões de linha e máscaras de coluna.

RPO

Veja objetivo de ponto de recuperação.

RTO

Veja objetivo de tempo de recuperação.

runbook

Um conjunto de procedimentos manuais ou automatizados necessários para realizar uma tarefa específica. Eles são normalmente criados para agilizar operações ou procedimentos repetitivos com altas taxas de erro.

S

SAML 2.0

Um padrão aberto que muitos provedores de identidade (IdPs) usam. Esse recurso permite o login único federado (SSO), para que os usuários possam fazer login Console de gerenciamento da AWS ou chamar as operações da AWS API sem que você precise criar um usuário no IAM para todos em sua organização. Para obter mais informações sobre a federação baseada em SAML 2.0, consulte Sobre a federação baseada em SAML 2.0 na documentação do IAM.

SCADA

Veja controle de supervisão e aquisição de dados.

SCP

Veja política de controle de serviço.

secret

Em AWS Secrets Manager, informações confidenciais ou restritas, como uma senha ou credenciais de usuário, que você armazena de forma criptografada. Consiste no valor secreto e em seus metadados. O valor secreto pode ser binário, uma única string ou várias strings. Para obter mais informações, consulte What's in a Secrets Manager secret? na documentação do Secrets Manager.

segurança desde a concepção

Uma abordagem em engenharia de sistemas que leva em consideração a segurança em todo o processo de desenvolvimento.

controle de segurança

Uma barreira de proteção técnica ou administrativa que impede, detecta ou reduz a capacidade de uma ameaça explorar uma vulnerabilidade de segurança. Existem quatro tipos primários de controles de segurança: preventivos, detectivos, responsivos e proativos.

hardening da segurança

O processo de reduzir a superfície de ataque para torná-la mais resistente a ataques. Isso pode incluir ações como remover recursos que não são mais necessários, implementar a prática recomendada de segurança de conceder privilégios mínimos ou desativar recursos desnecessários em arquivos de configuração.

sistema de gerenciamento de eventos e informações de segurança (SIEM)

Ferramentas e serviços que combinam sistemas de gerenciamento de informações de segurança (SIM) e gerenciamento de eventos de segurança (SEM). Um sistema SIEM coleta, monitora e analisa dados de servidores, redes, dispositivos e outras fontes para detectar ameaças e violações de segurança e gerar alertas.

automação de resposta de segurança

Uma ação predefinida e programada projetada para responder ou remediar automaticamente um evento de segurança. Essas automações servem como controles de segurança responsivos ou detectivos que ajudam você a implementar as melhores práticas AWS de segurança. Exemplos de ações de resposta automatizada incluem a modificação de um grupo de segurança da VPC, a aplicação de patches em uma instância do Amazon EC2 ou a alternância de credenciais.

Criptografia do lado do servidor

Criptografia dos dados em seu destino, por AWS service (Serviço da AWS) quem os recebe.

política de controle de serviços (SCP)

Uma política que fornece controle centralizado sobre as permissões de todas as contas em uma organização em AWS Organizations. SCPs defina barreiras ou estabeleça limites nas ações que um administrador pode delegar a usuários ou funções. Você pode usar SCPs como listas de permissão ou listas de negação para especificar quais serviços ou ações são permitidos ou proibidos. Para obter mais informações, consulte Políticas de controle de serviço na AWS Organizations documentação.

service endpoint (endpoint de serviço)

O URL do ponto de entrada para um AWS service (Serviço da AWS). Você pode usar o endpoint para se conectar programaticamente ao serviço de destino. Para obter mais informações, consulte Endpoints do AWS service (Serviço da AWS) na Referência geral da AWS.

acordo de serviço (SLA)

Um acordo que esclarece o que uma equipe de TI promete fornecer aos clientes, como tempo de atividade e performance do serviço.

indicador de nível de serviço (SLI)

Uma avaliação de um aspecto de performance de um serviço, como taxa de erro, disponibilidade ou throughput.

objetivo de nível de serviço (SLO)

Uma métrica alvo que representa a integridade de um serviço, conforme avaliado por um indicador de nível de serviço.

modelo de responsabilidade compartilhada

Um modelo que descreve a responsabilidade com a qual você compartilha AWS pela segurança e conformidade na nuvem. AWS é responsável pela segurança da nuvem, enquanto você é responsável pela segurança na nuvem. Para obter mais informações, consulte o Modelo de responsabilidade compartilhada.

SIEM

Veja sistema de gerenciamento de eventos e informações de segurança.

ponto único de falha (SPOF)

Uma falha em um único componente crítico de uma aplicação que pode interromper o sistema.

SLA

Veja cacordo de serviço.

SLI

Veja indicador de nível de serviço.

SLO

Veja objetivo de nível de serviço.

split-and-seed modelo

Um padrão para escalar e acelerar projetos de modernização. À medida que novos recursos e lançamentos de produtos são definidos, a equipe principal se divide para criar novas equipes de produtos. Isso ajuda a escalar os recursos e os serviços da sua organização, melhora a produtividade do desenvolvedor e possibilita inovações rápidas. Para obter mais informações, consulte Phased approach to modernizing applications in the Nuvem AWS.

SPOF

Veja ponto único de falha.

esquema em estrela

Uma estrutura organizacional de banco de dados que usa uma grande tabela de fatos para armazenar dados transacionais ou medidos e usa uma ou mais tabelas dimensionais menores para armazenar atributos de dados. Essa estrutura foi projetada para ser usada em um data warehouse ou para fins de inteligência comercial.

padrão strangler fig

Uma abordagem à modernização de sistemas monolíticos que consiste em reescrever e substituir incrementalmente a funcionalidade do sistema até que o sistema herdado possa ser desativado. Esse padrão usa a analogia de uma videira que cresce e se torna uma árvore estabelecida e, eventualmente, supera e substitui sua hospedeira. O padrão foi apresentado por Martin Fowler como forma de gerenciar riscos ao reescrever sistemas monolíticos. Para ver um exemplo de como aplicar esse padrão, consulte Modernizar incrementalmente os serviços Web herdados do Microsoft ASP.NET (ASMX) usando contêineres e o Amazon API Gateway.

sub-rede

Um intervalo de endereços IP na VPC. Cada sub-rede fica alocada em uma única zona de disponibilidade.

controle supervisório e aquisição de dados (SCADA)

Na manufatura, um sistema que usa hardware e software para monitorar ativos físicos e operações de produção.

symmetric encryption (criptografia simétrica)

Um algoritmo de criptografia que usa a mesma chave para criptografar e descriptografar dados.

testes sintéticos

Testar um sistema de forma que simule as interações do usuário para detectar possíveis problemas ou monitorar a performance. Você pode usar o Amazon CloudWatch Synthetics para criar esses testes.

prompt do sistema

Uma técnica para fornecer contexto, instruções ou orientações a um LLM a fim de direcionar seu comportamento. Os prompts do sistema ajudam a definir o contexto e a estabelecer regras para interações com os usuários.

T

tags

Pares de valores-chave que atuam como metadados para organizar seus recursos. AWS As tags podem ajudar você a gerenciar, identificar, organizar, pesquisar e filtrar recursos da . Para obter mais informações, consulte Marcar seus recursos do AWS.

variável-alvo

O valor que você está tentando prever no ML supervisionado. Ela também é conhecida como variável de resultado. Por exemplo, em uma configuração de fabricação, a variável-alvo pode ser um defeito do produto.

lista de tarefas

Uma ferramenta usada para monitorar o progresso por meio de um runbook. Uma lista de tarefas contém uma visão geral do runbook e uma lista de tarefas gerais a serem concluídas. Para cada tarefa geral, ela inclui o tempo estimado necessário, o proprietário e o progresso.

ambiente de teste

Veja ambiente.

treinamento

O processo de fornecer dados para que seu modelo de ML aprenda. Os dados de treinamento devem conter a resposta correta. O algoritmo de aprendizado descobre padrões nos dados de treinamento que mapeiam os atributos dos dados de entrada no destino (a resposta que você deseja prever). Ele gera um modelo de ML que captura esses padrões. Você pode usar o modelo de ML para obter previsões de novos dados cujo destino você não conhece.

gateway de trânsito

Um hub de trânsito de rede que você pode usar para interconectar sua rede com VPCs a rede local. Para obter mais informações, consulte O que é um gateway de trânsito na AWS Transit Gateway documentação.

fluxo de trabalho baseado em troncos

Uma abordagem na qual os desenvolvedores criam e testam recursos localmente em uma ramificação de recursos e, em seguida, mesclam essas alterações na ramificação principal. A ramificação principal é então criada para os ambientes de desenvolvimento, pré-produção e produção, sequencialmente.

Acesso confiável

Conceder permissões a um serviço que você especifica para realizar tarefas em sua organização AWS Organizations e em suas contas em seu nome. O serviço confiável cria um perfil vinculado ao serviço em cada conta, quando esse perfil é necessário, para realizar tarefas de gerenciamento para você. Para obter mais informações, consulte Usando AWS Organizations com outros AWS serviços na AWS Organizations documentação.

tuning (ajustar)

Alterar aspectos do processo de treinamento para melhorar a precisão do modelo de ML. Por exemplo, você pode treinar o modelo de ML gerando um conjunto de rótulos, adicionando rótulos e repetindo essas etapas várias vezes em configurações diferentes para otimizar o modelo.

equipe de duas pizzas

Uma pequena DevOps equipe que você pode alimentar com duas pizzas. Uma equipe de duas pizzas garante a melhor oportunidade possível de colaboração no desenvolvimento de software.

U

incerteza

Um conceito que se refere a informações imprecisas, incompletas ou desconhecidas que podem minar a confiabilidade dos modelos preditivos de ML. Há dois tipos de incertezas: a incerteza epistêmica é causada por dados limitados e incompletos, enquanto a incerteza aleatória é causada pelo ruído e pela aleatoriedade inerentes aos dados. Para obter mais informações, consulte o guia Como quantificar a incerteza em sistemas de aprendizado profundo.

tarefas indiferenciadas

Também conhecido como trabalho pesado, trabalho necessário para criar e operar um aplicativo, mas que não fornece valor direto ao usuário final nem oferece vantagem competitiva. Exemplos de tarefas indiferenciadas incluem aquisição, manutenção e planejamento de capacidade.

ambientes superiores

Veja ambiente.

V

aspiração

Uma operação de manutenção de banco de dados que envolve limpeza após atualizações incrementais para recuperar armazenamento e melhorar a performance.

controle de versões

Processos e ferramentas que rastreiam mudanças, como alterações no código-fonte em um repositório.

emparelhamento da VPC

Uma conexão entre duas VPCs que permite rotear o tráfego usando endereços IP privados. Para ter mais informações, consulte O que é emparelhamento de VPC? na documentação da Amazon VPC.

Vulnerabilidade

Uma falha de software ou hardware que compromete a segurança do sistema.

W

cache quente

Um cache de buffer que contém dados atuais e relevantes que são acessados com frequência. A instância do banco de dados pode ler do cache do buffer, o que é mais rápido do que ler da memória principal ou do disco.

dados mornos

Dados acessados raramente. Ao consultar esse tipo de dados, consultas moderadamente lentas geralmente são aceitáveis.

função de janela

Uma função SQL que executa um cálculo em um grupo de linhas que se relacionam de alguma forma com o registro atual. As funções de janela são úteis para processar tarefas, como calcular uma média móvel ou acessar o valor das linhas com base na posição relativa da linha atual.

workload

Uma coleção de códigos e recursos que geram valor empresarial, como uma aplicação voltada para o cliente ou um processo de backend.

workstreams

Grupos funcionais em um projeto de migração que são responsáveis por um conjunto específico de tarefas. Cada workstream é independente, mas oferece suporte aos outros workstreams do projeto. Por exemplo, o workstream de portfólio é responsável por priorizar aplicações, planejar ondas e coletar metadados de migração. O workstream de portfólio entrega esses ativos ao workstream de migração, que então migra os servidores e as aplicações.

WORM

Veja gravação única e várias leituras.

WQF

Veja AWS Workload Qualification Framework.

gravação única e várias leituras (WORM)

Um modelo de armazenamento que grava dados uma única vez e evita que os dados sejam excluídos ou modificados. Os usuários autorizados podem ler os dados quantas vezes forem necessárias, mas não podem alterá-los. Essa infraestrutura de armazenamento de dados é considerada imutável.

Z

exploração de dia zero

Um ataque, normalmente malware, que tira proveito de uma vulnerabilidade zero-day.

vulnerabilidade de dia zero

Uma falha ou vulnerabilidade não mitigada em um sistema de produção. Os agentes de ameaças podem usar esse tipo de vulnerabilidade para atacar o sistema. Os desenvolvedores frequentemente ficam cientes da vulnerabilidade como resultado do ataque.

prompt zero shot

Fornecer a um LLM instruções para realizar uma tarefa, mas sem exemplos (shots) que possam ajudar a orientá-lo. O LLM deve usar seu conhecimento pré-treinado para lidar com a tarefa. A

eficácia dos prompts zero-shot depende da complexidade da tarefa e da qualidade do prompt. Veja também prompts few-shot.

aplicação zumbi

Uma aplicação que tem um uso médio de CPU e memória inferior a 5%. Em um projeto de migração, é comum retirar essas aplicações.

As traduções são geradas por tradução automática. Em caso de conflito entre o conteúdo da tradução e da versão original em inglês, a versão em inglês prevalecerá.