Construindo um sistema escalável de rastreamento da web para dados do ESG em AWS

AWS Orientação prescritiva

AWS Orientação prescritiva: Construindo um sistema escalável de rastreamento da web para dados do ESG em AWS

Table of Contents

Construindo um sistema escalável de rastreamento da web para dados do ESG em AWS

Vijit Vashishtha e Mansi Doshi, da Amazon Web Services

Janeiro de 2025 (histórico do documento)

Fatores ambientais, sociais e de governança (ESG) são considerações críticas para os investidores ao avaliar possíveis investimentos:

- Ambiental — Concentra-se no impacto de uma empresa no mundo natural. Inclui fatores como emissões de carbono, gerenciamento de recursos e eficiência energética.
- Social — examina como uma empresa gerencia relacionamentos com funcionários, fornecedores, clientes e comunidades. Abrange aspectos como práticas trabalhistas, diversidade e envolvimento da comunidade.
- Governança — analisa a liderança, os controles internos e os direitos dos acionistas de uma empresa. Inclui composição do conselho, remuneração de executivos e ética nos negócios.

As empresas com práticas robustas de ESG são cada vez mais vistas como melhor posicionadas para sustentabilidade e lucratividade a longo prazo. Há uma demanda crescente dos investidores por informações ESG. As empresas que conseguem demonstrar suas credenciais de sustentabilidade por meio de dados ESG confiáveis e úteis estão melhor posicionadas para atrair capital e permanecer competitivas. As empresas publicam dados ESG por meio de várias fontes, como notícias, artigos e relatórios anuais. Como essas informações estão dispersas, um rastreador da Web pode ajudá-lo a coletar esses dados com eficiência.

Este guia abrangente demonstra como usar AWS Fargateo Amazon Elastic Compute Cloud (Amazon EC2) e o Amazon Simple Storage Service (Amazon S3) para criar um pipeline de coleta de dados robusto, escalável e responsável. AWS Batch Ele discute o seguinte:

- Arquitetando um sistema de rastreamento escalável usando o seguinte: Serviços da AWS
  - Fargate ou Amazon EC2 para executar o aplicativo de rastreamento
  - AWS Batch para orquestrar com eficiência trabalhos de rastreamento em grande escala
  - Amazon S3 para armazenamento de dados seguro e durável
- Implementando as melhores práticas de rastreamento ético, incluindo:

  - Respeitando o robots.txt e as políticas do site
  - Gerenciando a limitação de taxa para evitar sobrecarregar os sites de destino
  - Garantir a privacidade dos dados e o uso responsável das informações coletadas
- Desenvolvendo um Pythonrastreador baseado que é otimizado para infraestrutura AWS
- Otimizando o desempenho do rastreador enquanto mantém os padrões éticos

Público-alvo

Este guia é destinado a engenheiros de dados e arquitetos de nuvem que desejam coletar com eficiência grandes quantidades de dados up-to-date ESG de sites públicos. É particularmente relevante para projetos que envolvem análise de mercado, avaliação financeira sustentável ou pesquisa financeira.

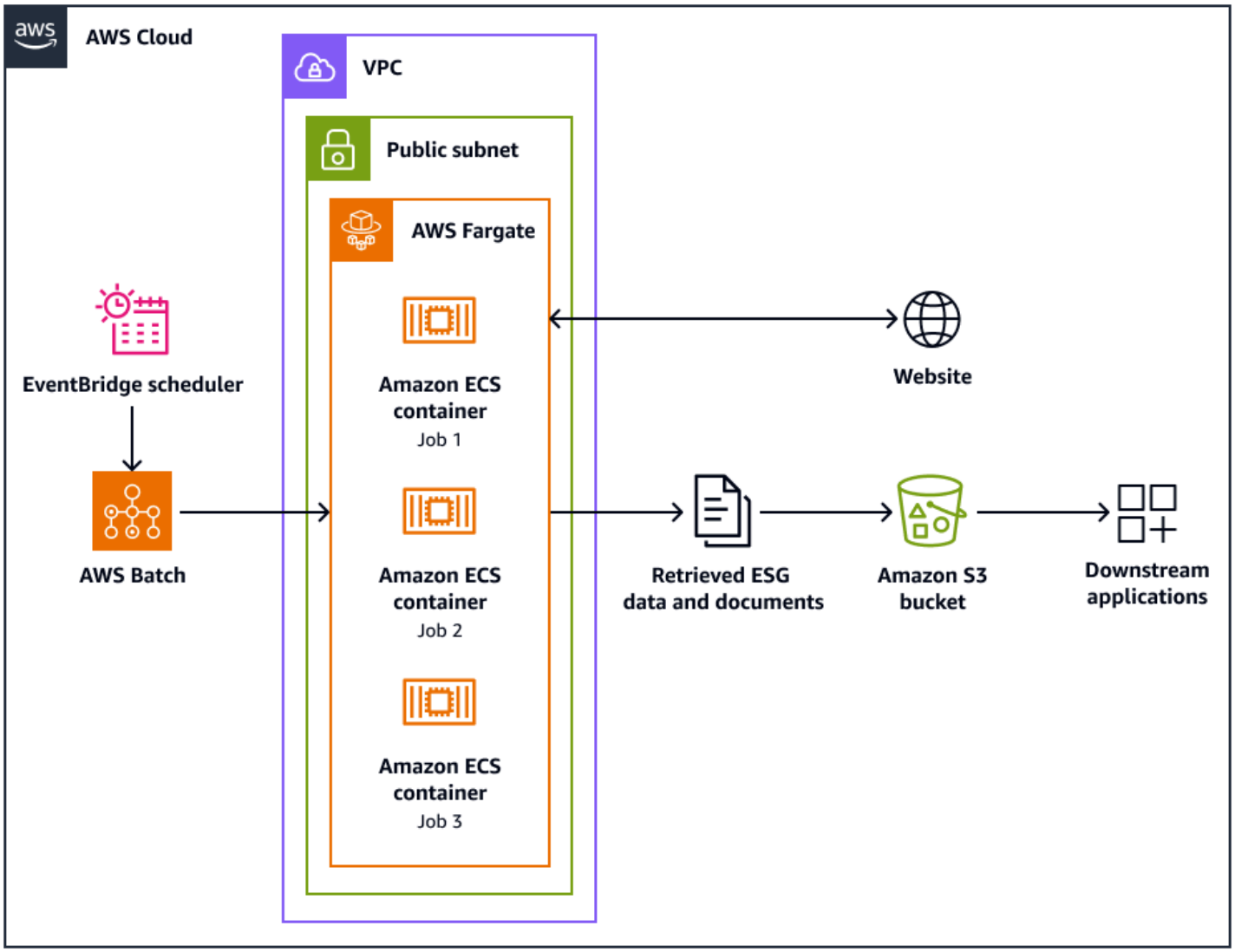
Resultados de negócios desejados

A seguir estão os motivos comuns pelos quais as empresas usam dados ESG:

- Gerenciamento de riscos — os dados do ESG ajudam você a identificar e mitigar riscos potenciais relacionados a questões ambientais, sociais e de governança.
- Atração de investidores — Muitos investidores agora consideram os fatores ESG ao tomar decisões de investimento. Eles veem as fortes práticas de ESG como indicadores de sustentabilidade e lucratividade a longo prazo.
- Gerenciamento de reputação — Um bom desempenho de ESG pode melhorar a reputação de uma empresa entre clientes, funcionários e o público em geral.
- Conformidade regulatória — À medida que as regulamentações relacionadas ao ESG aumentam, a adoção de práticas de ESG ajuda as empresas a se manterem à frente dos requisitos de conformidade.
- Inovação e eficiência — O foco nos fatores ESG pode impulsionar a inovação em produtos, serviços e operações. Isso leva a uma maior eficiência e economia de custos.
- Vantagem competitiva — O forte desempenho do ESG pode diferenciar uma empresa de seus concorrentes e abrir novas oportunidades de mercado.
- Engajamento das partes interessadas — As práticas de ESG ajudam as empresas a se engajar melhor e atender às expectativas de várias partes interessadas, incluindo funcionários, clientes e comunidades locais.

Arquitetura para um sistema escalável de rastreamento da web em AWS

O diagrama de arquitetura a seguir mostra um sistema de rastreador da Web projetado para extrair de forma ética dados ambientais, sociais e de governança (ESG) dos sites. Você usa um Pythonrastreador baseado que é otimizado para AWS infraestrutura. Você usa AWS Batch para orquestrar os trabalhos de rastreamento em grande escala e usa o Amazon Simple Storage Service (Amazon S3) para armazenamento. Os aplicativos downstream podem ingerir e armazenar os dados do bucket do Amazon S3.



O diagrama mostra o seguinte fluxo de trabalho:

1. O Amazon EventBridge Scheduler inicia o processo de rastreamento em um intervalo programado por você.
2. AWS Batch gerencia a execução dos trabalhos do rastreador da web. A fila de AWS Batch trabalhos retém e orquestra os trabalhos de rastreamento pendentes.
3. Os trabalhos de rastreamento da web são executados em contêineres do Amazon Elastic Container Service (Amazon ECS) em. AWS Fargate Os trabalhos são executados em uma sub-rede pública de uma nuvem privada virtual (VPC).
4. O rastreador da web rastreia o site de destino e recupera dados e documentos do ESG, como PDF, CSV ou outros arquivos de documentos.
5. O rastreador da web armazena os dados recuperados e os arquivos brutos em um bucket do Amazon S3.
6. Outros sistemas ou aplicativos ingerem ou processam os dados e arquivos armazenados no bucket do Amazon S3.

Design e operações de rastreadores da Web

Alguns sites são projetados especificamente para serem executados em desktops ou dispositivos móveis. O rastreador da Web foi projetado para oferecer suporte ao uso de um agente de usuário de desktop ou de um agente de usuário móvel. Esses agentes ajudam você a fazer solicitações com sucesso ao site de destino.

Depois que o rastreador da web é inicializado, ele executa as seguintes operações:

1. O rastreador da web chama o `setup()` método. Esse método busca e analisa o arquivo robots.txt.

   > **Note**
   >
   > Você também pode configurar o rastreador da web para buscar e analisar o mapa do site.

2. O rastreador da web processa o arquivo robots.txt. Se um atraso de rastreamento for especificado no arquivo robots.txt, o rastreador da Web extrairá o atraso de rastreamento para o agente de usuário de desktop. Se um atraso de rastreamento não for especificado no arquivo robots.txt, o rastreador da Web usará um atraso aleatório.
3. O rastreador da web chama o `crawl()` método, que inicia o processo de rastreamento. Se não URLs houver nenhum na fila, ele adiciona a URL inicial.

> **Note**
>
> O rastreador continua até atingir o número máximo de páginas ou ficar sem espaço URLs para rastrear.

4. O rastreador processa o. URLs Para cada URL na fila, o rastreador verifica se a URL já foi rastreada.
5. Se um URL não tiver sido rastreado, o rastreador chama o método da `crawl_url()` seguinte forma:
   a. O rastreador verifica o arquivo robots.txt para determinar se ele pode usar o agente de usuário de desktop para rastrear a URL.
   b. Se permitido, o rastreador tentará rastrear o URL usando o agente de usuário de desktop.
   c. Se não for permitido ou se o user agent de desktop falhar no rastreamento, o rastreador verificará o arquivo robots.txt para determinar se ele pode usar o agente de usuário móvel para rastrear o URL.
   d. Se permitido, o rastreador tentará rastrear o URL usando o agente de usuário móvel.
6. O rastreador chama o `attempt_crawl()` método, que recupera e processa o conteúdo. O rastreador envia uma solicitação GET para o URL com os cabeçalhos apropriados. Se a solicitação falhar, o rastreador usa a lógica de repetição.
7. Se o arquivo estiver no formato HTML, o rastreador chamará o `extract_esg_data()` método. Ele usa Beautiful Souppara analisar o conteúdo HTML. Ele extrai dados ambientais, sociais e de governança (ESG) usando a correspondência de palavras-chave.

   Se o arquivo for um PDF, o rastreador chamará o `save_pdf()` método. O rastreador baixa e salva o arquivo PDF no bucket do Amazon S3.
8. O rastreador chama o `extract_news_links()` método. Isso encontra e armazena links para artigos de notícias, comunicados à imprensa e postagens de blog.
9. O rastreador chama o `extract_pdf_links()` método. Isso identifica e armazena links para documentos PDF.
10. O rastreador chama o `is_relevant_to_sustainable_finance()` método. Isso verifica se as notícias ou artigos estão relacionados a finanças sustentáveis usando palavras-chave predefinidas.

11. Depois de cada tentativa de rastreamento, o rastreador implementa um atraso usando o método. `delay()` Se um atraso foi especificado no arquivo robots.txt, ele usa esse valor. Caso contrário, ele usa um atraso aleatório entre 1 e 3 segundos.
12. O rastreador chama o `save_esg_data()` método para salvar os dados do ESG em um arquivo CSV. O arquivo CSV é salvo no bucket do Amazon S3.
13. O rastreador chama o `save_news_links()` método para salvar os links de notícias em um arquivo CSV, incluindo informações relevantes. O arquivo CSV é salvo no bucket do Amazon S3.
14. O rastreador chama o `save_pdf_links()` método para salvar os links do PDF em um arquivo CSV. O arquivo CSV é salvo no bucket do Amazon S3.

Tratamento em lotes e processamento de dados

O processo de rastreamento é organizado e executado de forma estruturada. AWS Batch atribui os trabalhos de cada empresa para que sejam executados paralelamente, em lotes. Cada lote se concentra no domínio e nos subdomínios de uma única empresa, conforme você os identificou em seu conjunto de dados. No entanto, os trabalhos no mesmo lote são executados sequencialmente para que não inundem o site com muitas solicitações. Isso ajuda o aplicativo a gerenciar a carga de trabalho de rastreamento com mais eficiência e a garantir que todos os dados relevantes sejam capturados para cada empresa.

Ao organizar o rastreamento da web em lotes específicos da empresa, isso conteineriza os dados coletados. Isso ajuda a evitar que os dados de uma empresa sejam misturados com dados de outras empresas.

O processamento em lotes ajuda o aplicativo a coletar dados da web de forma eficiente, mantendo uma estrutura clara e uma separação de informações com base nas empresas-alvo e seus respectivos domínios da web. Essa abordagem ajuda a garantir a integridade e a usabilidade dos dados coletados, pois estão bem organizados e associados à empresa e aos domínios apropriados.

Construindo um sistema escalável de rastreamento da web em AWS

Esta seção descreve como criar o rastreador da Web descrito na Arquitetura seção. Ele inclui uma abordagem sistemática para criar um conjunto de dados robusto de empresas e suas propriedades web associadas. Esse conjunto de dados serve como base para suas atividades de rastreamento. Em seguida, esta seção descreve como criar um rastreador da Web ético em Python.

Tópicos

Preparando um conjunto de dados

Se você ainda não tiver feito isso, prepare um conjunto de dados detalhado dos sites dos quais você deseja coletar informações. Esse conjunto de dados deve incluir nomes de domínio do URL do site e nomes de subdomínios relevantes. Esta seção fornece um step-by-step processo para criar esse conjunto de dados.

Para preparar um conjunto de dados

1. Defina o escopo — determine o setor ou os setores em que você está se concentrando. Decida quantas empresas incluir. E defina todos os critérios que você deseja coletar sobre essas empresas, como número de funcionários, localização ou receita.
2. Identificar fontes de dados — Identifique quais fontes de informação você pode usar para coletar informações sobre essas empresas. Os exemplos incluem diretórios comerciais (como Crunchbase, Bloomberg ou Forbes), bolsas de valores (como NYSE e NASDAQ), associações ou publicações específicas do setor ou bancos de dados governamentais (como registros da SEC).
3. Crie uma tabela — Em sua ferramenta preferida, como Microsoft Excel, Planilhas Google ou um sistema de gerenciamento de banco de dados, crie uma tabela para coletar critérios sobre cada empresa. Inclua uma coluna para cada critério. No mínimo, inclua colunas para o nome da empresa, domínio principal, subdomínios, setor, tamanho e localização.

4. Colete as informações iniciais da empresa — Colete as seguintes informações sobre cada empresa e insira-as na tabela que você criou:
   - Company name (Nome da empresa)
   - Indústria ou setor
   - Porte da empresa (número de funcionários)
   - Revenue (Receita)
   - Localização da sede da empresa
5. Colete informações de domínio — Para cada empresa, extraia o nome de domínio principal do URL do site principal, como`example.com`. Você pode verificar as informações do domínio usando uma ferramenta de pesquisa de domínio WHOIS.
6. Reúna informações do subdomínio — Para cada empresa, pesquise os subdomínios registrados, como. `blog.example.com` Você pode usar ferramentas de enumeração de subdomínios, como Sublist3r, OWASP Amass ou Subfinder. Você pode executar o Google dorking (pesquisando`site:example.com`), verificar os registros DNS usando um `dig` comando ou uma ferramenta de pesquisa de DNS ou analisar certificados SSL ou TLS.
7. Valide e limpe os dados — revise, verifique e padronize os dados que você coletou. Por exemplo, remova todas as entradas duplicadas, remova informações de URL desnecessárias de domínios e subdomínios e verifique se todos os domínios e subdomínios estão ativos.
8. (Opcional) Categorize os subdomínios — Você pode categorizar os subdomínios em tipos. A seguir estão alguns exemplos de categorias que você pode encontrar:
   - Blogs, como `blog.example.com`
   - Support ou ajuda, como `support.example.com` ou `help.example.com`
   - Comércio eletrônico, como `shop.example.com` ou `store.example.com`
   - Recursos para desenvolvedores, como `dev.example.com` ou `api.example.com`
   - Regiões ou locais, como `us.example.com` ou `uk.example.com`
9. (Opcional) Adicionar metadados relevantes — Você pode registrar qualquer metadado relevante no conjunto de dados. Por exemplo, você pode adicionar a data da última atualização, a fonte das informações ou sua pontuação de confiança para a precisão do subdomínio.
10. Implemente o controle de versão — Use um sistema de controle de versão, como o Git, para acompanhar as alterações na tabela ao longo do tempo. Faça backup do conjunto de dados regularmente.

11. Manter a tabela — Configure um cronograma, como trimestral, para atualizar a tabela. Padronize e implemente um processo para adicionar novas empresas ou remover aquelas que você não precisa mais. Quando possível, automatize a descoberta de subdomínios.

Construindo o rastreador da web

Conforme descrito na Arquitetura seção, o aplicativo é executado em lotes — um para cada empresa.

Tópicos

- Capturando e processando o arquivo robots.txt
- Capturando e processando o mapa do site
- Projetando o rastreador

Capturando e processando o arquivo robots.txt

Depois de preparar o conjunto de dados, você precisa confirmar se o domínio tem um arquivo robots.txt. Para rastreadores da web e outros bots, o arquivo robots.txt indica quais seções do site eles podem visitar. Cumprir as instruções desse arquivo é uma prática recomendada importante para rastrear sites de forma ética. Para obter mais informações, consulte Práticas recomendadas para rastreadores da Web éticos neste guia.

Para capturar e processar o arquivo robots.txt

1. Se você ainda não tiver feito isso, instale a `requests` biblioteca executando o seguinte comando em um terminal:

```
pip install requests
```

2. Consulte o script a seguir. Este script faz o seguinte:

   - Ele define uma `check_robots_txt` função que usa um domínio como entrada.
   - Ele constrói o URL completo para o arquivo robots.txt.
   - Ele envia uma solicitação GET para a URL do arquivo robots.txt.
   - Se a solicitação for bem-sucedida (código de status 200), existirá um arquivo robots.txt.

- Se a solicitação falhar ou retornar um código de status diferente, o arquivo robots.txt não existe ou não está acessível.

```
import requests
from urllib.parse import urljoin
def check_robots_txt(domain):
    # Ensure the domain starts with a protocol
    if not domain.startswith(('http://', 'https://')):
        domain = 'https://' + domain
    # Construct the full URL for robots.txt
    robots_url = urljoin(domain, '/robots.txt')
    try:
        # Send a GET request to the robots.txt URL
        response = requests.get(robots_url, timeout=5)
        # Check if the request was successful (status code 200)
        if response.status_code == 200:
            print(f"robots.txt found at {robots_url}")
            return True
        else:
            print(f"No robots.txt found at {robots_url} (Status code:
 {response.status_code})")
            return False
    except requests.RequestException as e:
        print(f"Error checking {robots_url}: {e}")
        return False
```

> **Note**
>
> Esse script trata exceções para erros de rede ou outros problemas.

3. Se existir um arquivo robots.txt, use o script a seguir para baixá-lo:

```
import requests

def download(self, url):
    response = requests.get(url, headers=self.headers, timeout=5)
    response.raise_for_status()  # Raise an exception for non-2xx responses
    return response.text

def download_robots_txt(self):
```

```
    # Append '/robots.txt' to the URL to get the robots.txt file's URL
    robots_url = self.url.rstrip('/') + '/robots.txt'
    try:
        response = download(robots_url)
        return response
    except requests.exceptions.RequestException as e:
        print(f"Error downloading robots.txt: {e}, \nGenerating sitemap using
combinations...")
        return e
```

> **Note**
>
> Esses scripts podem ser personalizados ou modificados de acordo com seu caso de uso. Você também pode combinar esses scripts.

Capturando e processando o mapa do site

Em seguida, você precisa processar o mapa do site. Você pode usar o mapa do site para focar no rastreamento de páginas importantes. Isso melhora a eficiência do rastreamento. Para obter mais informações, consulte Práticas recomendadas para rastreadores da Web éticos neste guia.

Para capturar e processar o mapa do site

- Consulte o script a seguir. Esse script define uma `check_and_download_sitemap` função que:
  - Aceita um URL base, um URL opcional do sitemap do robots.txt e uma string de agente de usuário.
  - Verifica várias localizações potenciais do sitemap, incluindo a de robots.txt (se fornecida).
  - Tenta baixar o mapa do site de cada local.
  - Verifica se o conteúdo baixado está no formato XML.
  - Chama a `parse_sitemap` função para extrair URLs o. Esta função:
    - Analisa o conteúdo XML do mapa do site.
    - Lida com sitemaps regulares e arquivos de índice de sitemaps.
    - Busca recursivamente subsitemaps se um índice de sitemap for encontrado.

```
import requests
from urllib.parse import urljoin
import xml.etree.ElementTree as ET

def check_and_download_sitemap(base_url, robots_sitemap_url=None,
 user_agent='SitemapBot/1.0'):
    headers = {'User-Agent': user_agent}
    sitemap_locations = [robots_sitemap_url, urljoin(base_url, '/sitemap.xml'),
 urljoin(base_url, '/sitemap_index.xml'),
        urljoin(base_url, '/sitemap/'), urljoin(base_url, '/sitemap/sitemap.xml')]

    for sitemap_url in sitemap_locations:
        if not sitemap_url:
            continue
        print(f"Checking for sitemap at: {sitemap_url}")
        try:
            response = requests.get(sitemap_url, headers=headers, timeout=10)
            if response.status_code == 200:
                content_type = response.headers.get('Content-Type', '')
                if 'xml' in content_type:
                    print(f"Successfully downloaded sitemap from {sitemap_url}")
                    return parse_sitemap(response.text)
                else:
                    print(f"Found content at {sitemap_url}, but it's not XML.
 Content-Type: {content_type}")
        except requests.RequestException as e:
            print(f"Error downloading sitemap from {sitemap_url}: {e}")
    print("No sitemap found.")
    return []

def parse_sitemap(sitemap_content):
    urls = []
    try:
        root = ET.fromstring(sitemap_content)
        # Handle both sitemap and sitemapindex
        for loc in root.findall('.//{http://www.sitemaps.org/schemas/
sitemap/0.9}loc'):
            urls.append(loc.text)

        # If it's a sitemap index, recursively fetch each sitemap
        if root.tag.endswith('sitemapindex'):
```

```
            all_urls = []
            for url in urls:
                print(f"Fetching sub-sitemap: {url}")
                sub_sitemap_urls = check_and_download_sitemap(url)
                all_urls.extend(sub_sitemap_urls)
            return all_urls
    except ET.ParseError as e:
        print(f"Error parsing sitemap XML: {e}")
    return urls


if __name__ == "__main__":
    base_url = input("Enter the base URL of the website: ")
    robots_sitemap_url = input("Enter the sitemap URL from robots.txt (or press
 Enter if none): ").strip() or None
    urls = check_and_download_sitemap(base_url, robots_sitemap_url)
    print(f"Found {len(urls)} URLs in sitemap:")
    for url in urls[:5]:  # Print first 5 URLs as an example
        print(url)
    if len(urls) > 5:
        print("...")
```

Projetando o rastreador

Em seguida, você cria o rastreador da web. O rastreador foi projetado para seguir as melhores práticas descritas Práticas recomendadas para rastreadores da web éticos neste guia. Esta `EthicalCrawler` aula demonstra vários princípios fundamentais do rastreamento ético:

- Buscando e analisando o arquivo robots.txt — O rastreador busca o arquivo robots.txt para o site de destino.
- Respeitando as permissões de rastreamento — Antes de rastrear qualquer URL, o rastreador verifica se as regras no arquivo robots.txt permitem o rastreamento desse URL. Se um URL não for permitido, o rastreador o ignorará e passará para o próximo URL.
- Honrando o atraso de rastreamento — O rastreador verifica se há uma diretiva de atraso de rastreamento no arquivo robots.txt. Se um for especificado, o rastreador usa esse atraso entre as solicitações. Caso contrário, ele usa um atraso padrão.
- Identificação do agente do usuário — O rastreador usa uma string personalizada do agente do usuário para se identificar nos sites. Se necessário, os proprietários de sites podem definir regras específicas para restringir ou permitir seu rastreador.

- Tratamento de erros e degradação normal — Se o arquivo robots.txt não puder ser buscado ou analisado, o rastreador prosseguirá com regras padrão conservadoras. Ele lida com erros de rede e respostas HTTP que não sejam 200.
- Rastreamento limitado — Para evitar sobrecarregar o servidor, há um limite para quantas páginas podem ser rastreadas.

O script a seguir é um pseudocódigo que explica como o rastreador da web funciona:

```
import requests
from urllib.parse import urljoin, urlparse
import time

class EthicalCrawler:
    def __init__(self, start_url, user_agent='EthicalBot/1.0'):
        self.start_url = start_url
        self.user_agent = user_agent
        self.domain = urlparse(start_url).netloc
        self.robots_parser = None
        self.crawl_delay = 1  # Default delay in seconds

    def can_fetch(self, url):
        if self.robots_parser:
            return self.robots_parser.allowed(url, self.user_agent)
        return True  # If no robots.txt, assume allowed but crawl conservatively

    def get_crawl_delay(self):
        if self.robots_parser:
            delay = self.robots_parser.agent(self.user_agent).delay
            if delay is not None:
                self.crawl_delay = delay
        print(f"Using crawl delay of {self.crawl_delay} seconds")

    def crawl(self, max_pages=10):
        self.get_crawl_delay()
        pages_crawled = 0
        urls_to_crawl = [self.start_url]
        while urls_to_crawl and pages_crawled < max_pages:
            url = urls_to_crawl.pop(0)
            if not self.can_fetch(url):
                print(f"robots.txt disallows crawling: {url}")
                continue
```

```
            try:
                response = requests.get(url, headers={'User-Agent': self.user_agent})
                if response.status_code == 200:
                    print(f"Successfully crawled: {url}")
                    # Here you would typically parse the content, extract links, etc.
                    # For this example, we'll just increment the counter
                    pages_crawled += 1
                else:
                    print(f"Failed to crawl {url}: HTTP {response.status_code}")
            except Exception as e:
                print(f"Error crawling {url}: {e}")

            # Respect the crawl delay
            time.sleep(self.crawl_delay)

        print(f"Crawling complete. Crawled {pages_crawled} pages.")
```

Para criar um rastreador web ético e avançado que colete dados do ESG

1. Copie o seguinte exemplo de código para o rastreador ético avançado da Web usado neste sistema:

```
import requests
from urllib.parse import urljoin, urlparse
import time
from collections import deque
import random
from bs4 import BeautifulSoup
import re
import csv
import os


class EnhancedESGCrawler:
    def __init__(self, start_url):
        self.start_url = start_url
        self.domain = urlparse(start_url).netloc
        self.desktop_user_agent = 'ESGEthicalBot/1.0'
        self.mobile_user_agent = 'Mozilla/5.0 (iPhone; CPU iPhone OS 14_0 like
 Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/14.0 Mobile/15E148
 Safari/604.1'
        self.robots_parser = None
```

```
        self.crawl_delay = None
        self.urls_to_crawl = deque()
        self.crawled_urls = set()
        self.max_retries = 2
        self.session = requests.Session()
        self.esg_data = []
        self.news_links = []
        self.pdf_links = []

    def setup(self):
        self.fetch_robots_txt() # Provided in Previous Snippet
        self.fetch_sitemap() # Provided in Previous Snippet

    def can_fetch(self, url, user_agent):
        if self.robots_parser:
            return self.robots_parser.allowed(url, user_agent)
        return True

    def delay(self):
        if self.crawl_delay is not None:
            time.sleep(self.crawl_delay)
        else:
            time.sleep(random.uniform(1, 3))

    def get_headers(self, user_agent):
        return {'User-Agent': user_agent,
                'Accept': 'text/html,application/xhtml+xml,application/
xml;q=0.9,image/webp,*/*;q=0.8',
                'Accept-Language': 'en-US,en;q=0.5', 'Accept-Encoding': 'gzip,
 deflate, br', 'DNT': '1',
                'Connection': 'keep-alive', 'Upgrade-Insecure-Requests': '1'}

    def extract_esg_data(self, url, html_content):
        soup = BeautifulSoup(html_content, 'html.parser')
        esg_data = {
            'url': url,
            'environmental': self.extract_environmental_data(soup),
            'social': self.extract_social_data(soup),
            'governance': self.extract_governance_data(soup)
        }
        self.esg_data.append(esg_data)
        # Extract news links and PDFs
        self.extract_news_links(soup, url)
```

```
        self.extract_pdf_links(soup, url)

    def extract_environmental_data(self, soup):
        keywords = ['carbon footprint', 'emissions', 'renewable energy', 'waste
management', 'climate change']
        return self.extract_keyword_data(soup, keywords)

    def extract_social_data(self, soup):
        keywords = ['diversity', 'inclusion', 'human rights', 'labor practices',
'community engagement']
        return self.extract_keyword_data(soup, keywords)

    def extract_governance_data(self, soup):
        keywords = ['board structure', 'executive compensation', 'shareholder
rights', 'ethics', 'transparency']
        return self.extract_keyword_data(soup, keywords)

    def extract_keyword_data(self, soup, keywords):
        text = soup.get_text().lower()
        return {keyword: len(re.findall(r'\b' + re.escape(keyword) + r'\b', text))
for keyword in keywords}

    def extract_news_links(self, soup, base_url):
        news_keywords = ['news', 'press release', 'article', 'blog',
'sustainability']
        for a in soup.find_all('a', href=True):
            if any(keyword in a.text.lower() for keyword in news_keywords):
                full_url = urljoin(base_url, a['href'])
                if full_url not in self.news_links:
                    self.news_links.append({'url': full_url, 'text':
a.text.strip()})

    def extract_pdf_links(self, soup, base_url):
        for a in soup.find_all('a', href=True):
            if a['href'].lower().endswith('.pdf'):
                full_url = urljoin(base_url, a['href'])
                if full_url not in self.pdf_links:
                    self.pdf_links.append({'url': full_url, 'text':
a.text.strip()})

    def is_relevant_to_sustainable_finance(self, text):
        keywords = ['sustainable finance', 'esg', 'green bond', 'social impact',
'environmental impact',
```

```
                    'climate risk', 'sustainability report', 'corporate
 responsibility']
        return any(keyword in text.lower() for keyword in keywords)

    def attempt_crawl(self, url, user_agent):
        for _ in range(self.max_retries):
            try:
                response = self.session.get(url,
 headers=self.get_headers(user_agent), timeout=10)
                if response.status_code == 200:
                    print(f"Successfully crawled: {url}")
                    if response.headers.get('Content-Type', '').startswith('text/
html'):
                        self.extract_esg_data(url, response.text)
                    elif response.headers.get('Content-Type',
 '').startswith('application/pdf'):
                        self.save_pdf(url, response.content)
                    return True
                else:
                    print(f"Failed to crawl {url}: HTTP {response.status_code}")
            except requests.RequestException as e:
                print(f"Error crawling {url} with {user_agent}: {e}")

            self.delay()
        return False

    def crawl_url(self, url):
        if not self.can_fetch(url, self.desktop_user_agent):
            print(f"Robots.txt disallows desktop user agent: {url}")
            if self.can_fetch(url, self.mobile_user_agent):
                print(f"Attempting with mobile user agent: {url}")
                return self.attempt_crawl(url, self.mobile_user_agent)
            else:
                print(f"Robots.txt disallows both user agents: {url}")
                return False

        return self.attempt_crawl(url, self.desktop_user_agent)

    def crawl(self, max_pages=100):
        self.setup()

        if not self.urls_to_crawl:
            self.urls_to_crawl.append(self.start_url)
```

```
        pages_crawled = 0
        while self.urls_to_crawl and pages_crawled < max_pages:
            url = self.urls_to_crawl.popleft()
            if url not in self.crawled_urls:
                if self.crawl_url(url):
                    pages_crawled += 1
                self.crawled_urls.add(url)
                self.delay()

        print(f"Crawling complete. Successfully crawled {pages_crawled} pages.")
        self.save_esg_data()
        self.save_news_links()
        self.save_pdf_links()

    def save_esg_data(self):
        with open('esg_data.csv', 'w', newline='', encoding='utf-8') as file:
            writer = csv.DictWriter(file, fieldnames=['url', 'environmental',
'social', 'governance'])
            writer.writeheader()
            for data in self.esg_data:
                writer.writerow({
                    'url': data['url'],
                    'environmental': ', '.join([f"{k}: {v}" for k, v in
data['environmental'].items()]),
                    'social': ', '.join([f"{k}: {v}" for k, v in
data['social'].items()]),
                    'governance': ', '.join([f"{k}: {v}" for k, v in
data['governance'].items()])
                })
        print("ESG data saved to esg_data.csv")

    def save_news_links(self):
        with open('news_links.csv', 'w', newline='', encoding='utf-8') as file:
            writer = csv.DictWriter(file, fieldnames=['url', 'text', 'relevant'])
            writer.writeheader()
            for news in self.news_links:
                writer.writerow({
                    'url': news['url'],
                    'text': news['text'],
                    'relevant':
self.is_relevant_to_sustainable_finance(news['text'])
                })
```

```
        print("News links saved to news_links.csv")

    def save_pdf_links(self):
        # Code for saving PDF in S3 or filesystem

    def save_pdf(self, url, content):
        # Code for saving PDF in S3 or filesystem

# Example usage
if __name__ == "__main__":
    start_url = input("Enter the starting URL to crawl for ESG data and news: ")
    crawler = EnhancedESGCrawler(start_url)
    crawler.crawl(max_pages=50)
```

2. Configure os vários atributos, incluindo agentes de usuário, coleções vazias e listas de armazenamento de dados. URLs
3. Ajuste as palavras-chave e os critérios de relevância no `is_relevant_to_sustainable_finance()` método para atender às suas necessidades específicas.
4. Verifique se o arquivo robots.txt permite rastrear o site e se você está usando o atraso de rastreamento e o agente de usuário especificados no arquivo robots.txt.
5. Considere fazer as seguintes personalizações no script de rastreador da web fornecido, conforme necessário para sua organização:
    - Implemente o `fetch_sitemap()` método para uma descoberta de URL mais eficiente.
    - Melhore o registro e o tratamento de erros para uso em produção.
    - Implemente uma análise de relevância de conteúdo mais sofisticada.
    - Adicione controles de profundidade e amplitude para limitar o escopo do rastreamento.

Construindo a AWS infraestrutura

Há muitos Serviços da AWS que você pode usar para criar a infraestrutura de rastreamento da web. A seção Arquitetura deste guia inclui uma solução proposta. Recomendamos que você considere usar o seguinte Serviços da AWS para criar a infraestrutura de suporte para seu rastreador da Web:

- Use a Amazon Virtual Private Cloud (Amazon VPC) para criar a VPC e as sub-redes.
- Inicie o processo de rastreamento usando o Amazon Scheduler. EventBridge

- Gerencie os trabalhos do rastreador da web usando trabalhos e filas de AWS Batchtrabalhos.
- Use uma das soluções a seguir para executar as tarefas do rastreador da web:
  - Contêineres do Amazon Elastic Container Service (Amazon ECS) em AWS Fargate
  - Instâncias do Amazon Elastic Compute Cloud (Amazon EC2)

    > **Note**
    >
    > Se seu aplicativo puder lidar com interrupções, considere usar Amazon EC2 Spot Instances por meio do Spot Fleet. Frotas de instâncias spot podem ajudar você a economizar significativamente nos custos de computação.

  - AWS Lambda funções
- Armazene os dados recuperados e os arquivos brutos em um bucket do Amazon Simple Storage Service (Amazon S3).

Práticas recomendadas para rastreadores da web éticos

Esta seção discute as melhores práticas e as principais considerações éticas para criar um aplicativo de rastreamento da web que coleta dados ambientais, sociais e de governança (ESG). Ao aderir a essas melhores práticas, você pode proteger seu projeto e sua organização e contribuir para um ecossistema web mais responsável e sustentável. Essa abordagem ajuda você a acessar dados valiosos e usá-los para pesquisa, negócios e inovação de uma forma que respeite todas as partes interessadas.

Conformidade com Robots.txt

O arquivo robots.txt é usado em sites para se comunicar com rastreadores e bots da Web sobre quais partes do site devem ou não ser acessadas ou rastreadas. Quando um rastreador da Web encontra um arquivo robots.txt em um site, ele analisa as instruções e ajusta seu comportamento de rastreamento adequadamente. Isso evita que o rastreador viole as instruções do proprietário do site e mantém um relacionamento cooperativo entre o site e o rastreador. Portanto, o arquivo robots.txt ajuda no controle de acesso, na proteção de conteúdo confidencial, no gerenciamento de carga e na conformidade legal.

Convém seguir estas práticas recomendadas:

- Sempre verifique e respeite as regras no arquivo robots.txt.
- Antes de rastrear qualquer URL, verifique as regras para agentes de usuário de desktop e dispositivos móveis.
- Se o site permitir somente agentes de usuários móveis, use um cabeçalho de agente diferente, como um cabeçalho de agente móvel, para sua solicitação.

A ausência de um arquivo robots.txt não significa necessariamente que você não pode ou não deve rastrear um site. O rastreamento deve sempre ser feito com responsabilidade, respeitando os recursos do site e os direitos implícitos do proprietário. Veja a seguir as melhores práticas recomendadas quando um robots.txt não está presente:

- Suponha que o rastreamento seja permitido, mas proceda com cuidado.
- Implemente práticas educadas de rastreamento.
- Considere entrar em contato com o proprietário do site para obter permissão se você planeja realizar um rastreamento extensivo.

Limitação da taxa de rastreamento

Use uma taxa de rastreamento razoável para evitar sobrecarregar o servidor. Implemente atrasos entre as solicitações, conforme especificado pelo arquivo robots.txt ou usando um atraso aleatório. Para sites pequenos ou médios, uma solicitação a cada 10 a 15 segundos pode ser apropriada. Para sites maiores ou com permissões explícitas de rastreamento, 1 a 2 solicitações por segundo podem ser adequadas.

Transparência entre usuário e agente

Identifique seu rastreador no cabeçalho do agente de usuário. Essas informações do cabeçalho HTTP têm como objetivo identificar o dispositivo que está solicitando o conteúdo. Normalmente, a palavra bot é incluída no nome do agente. Às vezes, rastreadores e outros bots usam um campo importante no cabeçalho para incluir informações de contato.

Rastreamento eficiente

Use o mapa do site, desenvolvido pelo proprietário do site, para se concentrar nas páginas importantes.

Abordagem adaptativa

Programe o rastreador para mudar para um agente de usuário móvel se a versão para desktop não for bem-sucedida. Isso pode fornecer acesso ao rastreador e reduzir a pressão sobre o servidor do site.

Gerenciamento de erros

Certifique-se de que o rastreador manipule vários códigos de status HTTP adequadamente. Por exemplo, o rastreador deve pausar se encontrar um código de status 429 (“Muitas solicitações”). Se o rastreador receber continuamente 403 códigos de status (“Proibido”), considere interromper o rastreamento.

Rastreamento em lotes

Recomendamos fazer o seguinte:

- Em vez de rastrear tudo de uma URLs vez, divida a tarefa em lotes menores. Isso pode ajudar a distribuir a carga e reduzir o risco de problemas, como tempos limite ou restrições de recursos.
- Se se espera que a tarefa geral de rastreamento seja de longa duração, considere dividi-la em várias tarefas menores e mais gerenciáveis. Isso pode tornar o processo mais escalável e resiliente.
- Se o número URLs de rastreamentos for relativamente pequeno, considere usar uma solução sem servidor, como. AWS Lambda As funções Lambda podem ser uma boa opção para tarefas de curta duração e orientadas por eventos, pois elas escalam e gerenciam automaticamente o gerenciamento de recursos.

Segurança

Para tarefas de computação de rastreamento na web, recomendamos que você configure o ambiente para permitir somente tráfego de saída. Isso ajuda a aumentar a segurança, minimizando a superfície de ataque e reduzindo o risco de acesso não autorizado de entrada. Permitir somente conexões de saída permite que o processo de rastreamento se comunique com os sites de destino e recupere os dados necessários, além de restringir qualquer tráfego de entrada que possa comprometer o sistema.

Outras considerações

Analise as seguintes considerações adicionais e as melhores práticas:

- Verifique as diretrizes de rastreamento nos termos de serviço ou na política de privacidade do site.
- Procure `meta` tags no HTML que possam fornecer diretivas de rastreamento.
- Esteja ciente das restrições legais em sua jurisdição em relação à coleta e uso de dados.
- Esteja preparado para parar de rastrear se solicitado pelo proprietário do site.

Perguntas frequentes

E se um arquivo robots.txt não estiver disponível?

A ausência de um arquivo robots.txt não significa necessariamente que você não pode ou não deve rastrear um site. O rastreamento deve sempre ser feito com responsabilidade, respeitando os recursos do site e os direitos implícitos do proprietário do site.

E se um arquivo sitemaps.xml não estiver disponível?

Dependendo do requisito, você pode fazer o seguinte:

- Pesquisar sitemaps em HTML — Procure uma página de sitemap em HTML que liste as páginas importantes do site. Eles geralmente estão vinculados no rodapé.
- Rastreamento a partir da página inicial — Comece a rastrear a partir da página inicial e siga os links internos para descobrir outras páginas.
- Analise padrões de URL — Analise a estrutura de URL do site para identificar padrões e gerar potencial de forma programática. URLs
- Examine o arquivo robots.txt — Verifique se há páginas ou diretórios não permitidos no arquivo robots.txt. Eles podem fornecer pistas sobre a estrutura do site.
- Analise os endpoints da API — Alguns sites oferecem endpoints de API que podem ser usados para recuperar conteúdo e estruturar informações.
- Verifique os resultados do mecanismo de pesquisa — Use os mecanismos de pesquisa para encontrar páginas indexadas do site usando o site: operador de pesquisa, como `site:example.com`.
- Analise backlinks — Analise os backlinks para o site para descobrir páginas importantes às quais outros sites estão vinculados.
- Analise os arquivos da web — Verifique os arquivos da Internet, como o Wayback Machine, para ver as versões mais antigas do site que poderiam ter mapas do site ou estruturas diferentes.
- Procure padrões do sistema de gerenciamento de conteúdo (CMS) — Se você puder identificar o CMS, use padrões de URL comuns associados a esse sistema.
- Confirme a JavaScript renderização — Se o site depender muito JavaScript, certifique-se de que seu rastreador possa renderizar JavaScript para descobrir conteúdo carregado dinamicamente.

Para alguns sites, o arquivo sitemap.xml é carregado depois que a JavaScript renderização é ativada.

Posso usar uma solução sem servidor em vez da Amazon ou do EC2 Amazon ECS?

Sim. AWS Lambdafunções de rastreamento na web podem ser uma opção viável, especialmente para tarefas de rastreamento em menor escala ou mais modulares. No entanto, para operações de rastreamento em grande escala e de longa duração, uma abordagem mais tradicional que usa instâncias do Amazon Elastic Compute Cloud (Amazon EC2) ou do Amazon Elastic Container Service (Amazon ECS) pode ser mais adequada. É importante avaliar cuidadosamente seus requisitos e desvantagens específicos ao escolher o serviço de computação certo para suas necessidades de rastreamento na web.

Por que o rastreador está recebendo um código de status 403?

HTTP 403 é um código de status HTTP que significa que o acesso ao recurso solicitado é proibido. Se a solicitação estiver correta, o servidor entendeu a solicitação e não a atenderá. Para evitar um código de status 403, você pode fazer o seguinte:

- Limite sua taxa de rastreamento.
- Verifique se o mapa do site ou o arquivo robots.txt permitem que o rastreador acesse o URL.
- Experimente com um agente de usuário móvel em vez de um agente de usuário para desktop.

Se nenhuma das opções acima funcionar, você deve respeitar a decisão dos proprietários do site e não rastrear a página.

Próximas etapas e recursos

Depois de coletar os dados ambientais, sociais e de governança (ESG) brutos, você pode fazer o seguinte para extrair informações significativas dos dados:

1. Limpe os dados — Os dados podem incluir uma quantidade significativa de informações irrelevantes que não estão relacionadas a fatores ESG e dados financeiros. É importante remover esses dados irrelevantes e reter somente as informações necessárias para realizar as análises necessárias. Você pode usar ferramentas, como yfinance, para ajudar a limpar os dados.
2. Extraia e transforme os dados — Extraia os recursos ou variáveis relevantes dos dados brutos e transforme-os em um formato adequado para análise. Você pode transformar os dados em formato tabular para melhor legibilidade e clareza. Você pode usar uma biblioteca, como pandas, para refinar os dados. Você também pode usar engenharia de recursos, normalização de dados e métricas derivadas para transformar os dados.
3. Realizar análises — Você pode realizar várias tarefas analíticas. Isso pode incluir a geração de estatísticas descritivas, a criação de visualizações de dados e a realização de análises exploratórias de dados para obter informações sobre o desempenho ESG das empresas.
4. Aplique o aprendizado de máquina — você pode usar os dados limpos e transformados para treinar modelos de aprendizado de máquina. Esses modelos podem ajudá-lo a identificar empresas que atualmente estão demonstrando sustentabilidade financeira e a projetar seu desempenho futuro em sustentabilidade.

Ao usar o rastreador da web e esse processo de avaliação de dados, você pode obter efetivamente uma compreensão abrangente das práticas de sustentabilidade e do desempenho financeiro das empresas que você está avaliando. Você pode usar essas informações para embasar decisões de investimento, acompanhar o progresso e apoiar práticas comerciais sustentáveis.

Recursos

- O que é um rastreador da web? (Cloudflare site)
- Guia de investimento em ESG (Investopedia site)

Ferramentas

- Beautiful Soup (Beautiful Soup documentação)
- Manipulação de dados tabulares e relacionais (pandas site)

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	—	6 de janeiro de 2025

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: migre seu banco de dados Oracle local para a edição compatível com o Amazon Aurora 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: Migre seu banco de dados Oracle local para o Amazon Relational Database Service (Amazon RDS) for Oracle no. Nuvem AWS
- Recomprar (drop and shop): mude para um produto diferente, normalmente migrando de uma licença tradicional para um modelo SaaS. Exemplo: migre seu sistema de gerenciamento de relacionamento com o cliente (CRM) para a 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: Migre seu banco de dados Oracle local para o Oracle em uma EC2 instância no. 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 local para um serviço em 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 atributos.

serviços abstratos

Veja os 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 no qual 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.

função agregada

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 a inteligência artificial.

AIOps

Veja as 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 aplicativos

Uma abordagem de segurança que permite o uso somente de aplicativos aprovados 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 ruim

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

BCP

Veja o 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 do aplicativo em um ambiente (azul) e a nova versão do aplicativo no outro ambiente (verde). Essa estratégia ajuda você a reverter rapidamente com o mínimo de impacto.

bot

Um aplicativo de software que executa tarefas automatizadas pela Internet e simula a atividade ou interação humana. Alguns bots são úteis ou benéficos, como rastreadores da Web que indexam informações na Internet. Alguns outros bots, conhecidos como bots ruins, têm como objetivo perturbar ou causar danos a indivíduos ou organizações.

botnet

Redes de bots infectadas por malware e sob o controle de uma única parte, conhecidas como pastor de bots ou operador de bots. As redes de bots 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 em vidro quebrado

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 Implementar procedimentos de quebra de vidro na orientação do Well-Architected AWS .

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

Consulte Estrutura de adoção da AWS nuvem.

implantação canária

O lançamento lento e incremental de uma versão para usuários finais. Quando estiver confiante, você implanta a nova versão e substituirá a versão atual em sua totalidade.

CCoE

Veja o Centro de Excelência em Nuvem.

CDC

Veja a captura de dados de alterações.

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 a integração e a 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 Empresarial.

computação em nuvem

A tecnologia de nuvem normalmente usada para armazenamento de dados remoto e gerenciamento de dispositivos de IoT. A computação em nuvem geralmente está conectada à tecnologia de computação de ponta.

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 quando migram para o 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

Consulte o 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 GitHub ouBitbucket 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 da IA que usa aprendizado de máquina para analisar e extrair informações de formatos visuais, como imagens e vídeos digitais. Por exemplo, a Amazon SageMaker AI fornece algoritmos de processamento de imagem para CV.

desvio de configuração

Para uma carga de trabalho, uma alteração de configuração em relação ao estado esperado. Isso pode fazer com que a carga de trabalho 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 is commonly described as a pipeline. CI/CDpode 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.

malha de dados

Uma estrutura arquitetônica 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 que oferece suporte à inteligência comercial, como análises. 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

Consulte a 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 o 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 são comumente 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 desastre. Para obter mais informações, consulte Recuperação de desastres de cargas de trabalho em AWS: Recuperação na nuvem no AWS Well-Architected Framework.

DML

Veja a 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 a recuperação de desastres.

detecção de deriva

Rastreando desvios de uma configuração básica. 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 o mapeamento do fluxo de valor do desenvolvimento.

E

EDA

Veja a 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 à computação em nuvem, a computação de ponta 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 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 o endpoint do 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 corporativos (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 pipeline de CI/CD, 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 o planejamento de recursos corporativos.

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. Ele armazena dados quantitativos sobre as operações comerciais. Normalmente, uma tabela de fatos contém dois tipos de colunas: aquelas que contêm medidas e aquelas que contêm uma chave externa para uma tabela de dimensões.

falham rapidamente

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

limite de isolamento de 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 Limites de isolamento de AWS falhas.

ramificação de recursos

Veja a filial.

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.

solicitação rápida

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 contextual, em que os modelos aprendem com exemplos (fotos) incorporados aos prompts. Solicitações rápidas podem ser eficazes para tarefas que exigem formatação, raciocínio ou conhecimento de domínio específicos. Veja também a solicitação zero-shot.

FGAC

Veja o 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 por meio da captura de dados alterados para migrar 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 o modelo da fundação.

modelo de fundação (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 básicos.

G

IA generativa

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

bloqueio geográfico

Veja as 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 baseado em troncos é a abordagem moderna e preferida.

imagem dourada

Um instantâneo de um sistema ou software usado como modelo para implantar novas instâncias desse sistema ou software. Por exemplo, na manufatura, uma imagem dourada 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, Amazon GuardDuty AWS Trusted Advisor, Amazon Inspector e verificações personalizadas AWS Lambda .

H

HA

Veja a 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 retenção

Uma parte dos dados históricos rotulados que são retidos de um conjunto de dados usado para treinar um modelo de aprendizado de máquina. Você pode usar dados de retenção para avaliar o desempenho do modelo comparando as previsões 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 típico de uma DevOps versão.

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 a 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 a Internet das Coisas industrial.

infraestrutura imutável

Um modelo que implanta uma nova infraestrutura para cargas de trabalho de produção em vez de atualizar, corrigir ou modificar a infraestrutura existente. Infraestruturas imutáveis são inerentemente mais consistentes, confiáveis e previsíveis do que infraestruturas mutáveis. Para obter mais informações, consulte as melhores práticas de implantação usando infraestrutura imutável no Well-Architected AWS 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 fabricação por meio de avanços em conectividade, dados em tempo real, automação, análise 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

Consulte 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

Consulte a biblioteca de informações de TI.

ITSM

Veja o 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.

modelo de linguagem grande (LLM)

Um modelo de IA de aprendizado profundo que é pré-treinado em uma grande quantidade de dados. Um LLM pode realizar várias tarefas, como responder 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 etiquetas.

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 um modelo de linguagem grande.

ambientes inferiores

Veja o 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 a filial.

malware

Software projetado para comprometer a segurança ou a privacidade do computador. O malware pode interromper os sistemas do computador, vazar informações confidenciais ou obter acesso não autorizado. Exemplos de malware incluem vírus, worms, ransomware, cavalos de Tróia, 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 abstratos.

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

Consulte Migration Acceleration Program.

mecanismo

Um processo completo no qual 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 o 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 a Amazon EC2 com o AWS 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 o. 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 carga de trabalho para o. Nuvem AWS Para obter mais informações, consulte a entrada de 7 Rs neste glossário e consulte Mobilize sua organização para acelerar migrações em grande escala.

ML

Veja o aprendizado de máquina.

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 Estratégia para modernizar aplicativos no 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 Avaliação da prontidão para modernização de aplicativos no. 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.

MAPA

Consulte Avaliação do portfólio de migração.

MQTT

Consulte 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 cargas de trabalho 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 o controle de acesso de origem.

CARVALHO

Veja a identidade de acesso de origem.

OCM

Veja o gerenciamento de mudanças 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 a integração de operações.

OLA

Veja o contrato em 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

Consulte Comunicação de processo aberto — Arquitetura unificada.

Comunicação de processo aberto — Arquitetura unificada (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 melhores práticas 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 Well-Architected AWS Framework.

tecnologia operacional (OT)

Sistemas de hardware e software que funcionam com o ambiente físico para controlar operações, equipamentos e infraestrutura industriais. Na manufatura, a integração dos sistemas OT e de tecnologia da informação (TI) 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 a análise de prontidão operacional.

OT

Veja a 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 as 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

Consulte controlador lógico programável.

AMEIXA

Veja o gerenciamento do ciclo de vida do produto.

política

Um objeto que pode definir permissões (consulte a política baseada em identidade), especificar as condições de acesso (consulte a política baseada em recursos) ou definir as permissões máximas para todas as contas em uma organização em AWS Organizations (consulte a 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. Para obter mais informações, consulte Habilitar a persistência de dados em microsserviços.

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 `WHERE` cláusula.

pressão de predicados

Uma técnica de otimização de consulta 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 o desempenho 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 de engenharia de sistema 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 projetado para impedir a implantação de recursos não compatíveis. Esses controles examinam 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 o design, desenvolvimento e lançamento, passando pelo crescimento e maturidade, até o declínio e a remoção.

ambiente de produção

Veja o ambiente.

controlador lógico programável (PLC)

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

encadeamento imediato

Usando a 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 no qual outros microsserviços possam se inscrever. 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, responsável, consultado, informado (RACI).

RAG

Consulte Geração Aumentada de 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, responsável, consultado, informado (RACI).

RCAC

Veja o 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.

rearquiteta

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 mais informações, consulte Especificar o que Regiões da AWS sua conta pode usar.

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 um aplicativo de resistir ou se recuperar de interrupções. Alta disponibilidade e recuperação de desastres são considerações comuns ao planejar a resiliência no. Nuvem AWS Para obter mais informações, consulte Nuvem AWS Resiliência.

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.

aposentar-se

Veja 7 Rs.

Geração Aumentada de Recuperação (RAG)

Uma tecnologia de IA generativa na qual 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.

alternância

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

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 o objetivo do ponto de recuperação.

RTO

Veja o objetivo do 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 AWS Management Console 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 a política de controle de serviços.

secret

Em AWS Secrets Manager, informações confidenciais ou restritas, como uma senha ou credenciais de usuário, que você armazena de forma criptografada. Ele 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 O que há em um segredo do Secrets Manager? na documentação do Secrets Manager.

segurança por design

Uma abordagem de engenharia de sistemas que leva em conta 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 principais de controles de segurança: preventivos, detectivos, responsivos e proativos.

fortalecimento 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 correção de uma instância EC2 da Amazon ou a rotação 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 medida de um aspecto de desempenho de um serviço, como taxa de erro, disponibilidade ou taxa de transferência.

objetivo de nível de serviço (SLO)

Uma métrica alvo que representa a integridade de um serviço, conforme medida 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 informações de segurança e sistema de gerenciamento de eventos.

ponto único de falha (SPOF)

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

SLA

Veja o contrato de nível de serviço.

ESGUIO

Veja o indicador de nível de serviço.

SLO

Veja o objetivo do 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 Abordagem em fases para modernizar aplicativos no. Nuvem AWS

CUSPE

Veja um único ponto de falha.

esquema de estrelas

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 uso 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 de supervisão 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 o desempenho. Você pode usar o Amazon CloudWatch Synthetics para criar esses testes.

prompt do sistema

Uma técnica para fornecer contexto, instruções ou diretrizes a um LLM para direcionar seu comportamento. Os prompts do sistema ajudam a definir o contexto e 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. 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 o 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 o 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 back-end.

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.

MINHOCA

Veja escrever uma vez, ler muitas.

WQF

Consulte Estrutura de qualificação AWS da carga de trabalho.

escreva uma vez, leia muitas (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, geralmente malware, que tira proveito de uma vulnerabilidade de dia zero.

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.

aviso zero-shot

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

eficácia da solicitação zero depende da complexidade da tarefa e da qualidade da solicitação. Veja também a solicitação de algumas fotos.

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.