



Building a scalable web crawling system for ESG data on AWS

# AWS Prescriptive Guidance



# **AWS Prescriptive Guidance: Building a scalable web crawling system for ESG data on AWS**

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon's trademarks and trade dress may not be used in connection with any product or service that is not Amazon's, in any manner that is likely to cause confusion among customers, or in any manner that disparages or discredits Amazon. All other trademarks not owned by Amazon are the property of their respective owners, who may or may not be affiliated with, connected to, or sponsored by Amazon.

---

# Table of Contents

|  |           |
|--|-----------|
| <b>Introduction .....</b>  | <b>1</b>  |
| Intended audience .....  | 2         |
| Targeted business outcomes .....   | 2         |
| <b>Architecture .....</b>  | <b>3</b>  |
| Web crawler design and operations .....                                    | 4         |
| Batching and data processing .....   | 6         |
| <b>Building the system .....</b>   | <b>7</b>  |
| Preparing a dataset .....  | 7         |
| Building the web crawler .....   | 9         |
| Capturing and processing the robots.txt file .....                         | 9         |
| Capturing and processing the sitemap .....                                 | 11        |
| Designing the crawler .....  | 13        |
| Building the AWS infrastructure .....                                      | 20        |
| <b>Best practices .....</b>  | <b>21</b> |
| Robots.txt compliance .....  | 21        |
| Crawl-rate limiting .....  | 22        |
| User-agent transparency .....  | 22        |
| Efficient crawling .....   | 22        |
| Adaptive approach .....  | 22        |
| Error handling .....   | 22        |
| Crawling in batches .....  | 22        |
| Security .....   | 23        |
| Other considerations .....   | 23        |
| <b>FAQ .....</b>   | <b>24</b> |
| What if a robots.txt file is not available? .....                          | 24        |
| What if a sitemaps.xml file is not available? .....                        | 24        |
| Can I use a serverless solution instead of Amazon EC2 or Amazon ECS? ..... | 25        |
| Why is the crawler getting a 403 status code? .....                        | 25        |
| <b>Next steps and resources .....</b>                                      | <b>26</b> |
| Resources .....  | 26        |
| Tools .....  | 26        |
| <b>Document history .....</b>  | <b>28</b> |
| <b>Glossary .....</b>  | <b>29</b> |
| # .....  | 29        |

|         |    |
|---------|----|
| A ..... | 30 |
| B ..... | 33 |
| C ..... | 35 |
| D ..... | 38 |
| E ..... | 42 |
| F ..... | 44 |
| G ..... | 46 |
| H ..... | 47 |
| I ..... | 48 |
| L ..... | 50 |
| M ..... | 52 |
| O ..... | 56 |
| P ..... | 58 |
| Q ..... | 61 |
| R ..... | 61 |
| S ..... | 64 |
| T ..... | 68 |
| U ..... | 69 |
| V ..... | 70 |
| W ..... | 70 |
| Z ..... | 71 |

# Building a scalable web crawling system for ESG data on AWS

*Vijit Vashishtha and Mansi Doshi, Amazon Web Services*

January 2025 ([document history](#))

Environmental, social, and governance (ESG) factors are critical considerations for investors when evaluating potential investments:

- **Environmental** — Focuses on a company's impact on the natural world. It includes factors like carbon emissions, resource management, and energy efficiency.
- **Social** — Examines how a company manages relationships with employees, suppliers, customers, and communities. It covers aspects like labor practices, diversity, and community engagement.
- **Governance** — Looks at a company's leadership, internal controls, and shareholder rights. It includes board composition, executive compensation, and business ethics.

Companies with robust ESG practices are increasingly viewed as better positioned for long-term sustainability and profitability. There is growing investor demand for ESG information. Companies that can demonstrate their sustainability credentials through reliable, useful ESG data are better positioned to attract capital and remain competitive. Companies publish ESG data through various sources, such as news, articles, and annual reports. Because this information is scattered, a web crawler can help you efficiently gather this data.

This comprehensive guide demonstrates how to use [AWS Fargate](#), [Amazon Elastic Compute Cloud \(Amazon EC2\)](#), [AWS Batch](#), and [Amazon Simple Storage Service \(Amazon S3\)](#) to build a robust, scalable, and responsible data collection pipeline. It discusses the following:

- Architecting a scalable crawling system by using the following AWS services:
  - Fargate or Amazon EC2 for running the crawler application
  - AWS Batch for efficiently orchestrating large-scale crawling jobs
  - Amazon S3 for secure and durable data storage
- Implementing best practices for ethical crawling, including:
  - Respecting robots.txt and website policies
  - Managing rate limiting to avoid overwhelming target sites

- Ensuring data privacy and responsible use of collected information
- Developing a Python-based crawler that is optimized for AWS infrastructure
- Optimizing crawler performance while maintaining ethical standards

## Intended audience

This guide is intended for data engineers and cloud architects who want to efficiently collect large amounts of up-to-date ESG data from public websites. It is particularly relevant for projects that involve market analysis, sustainable financial assessment, or financial research.

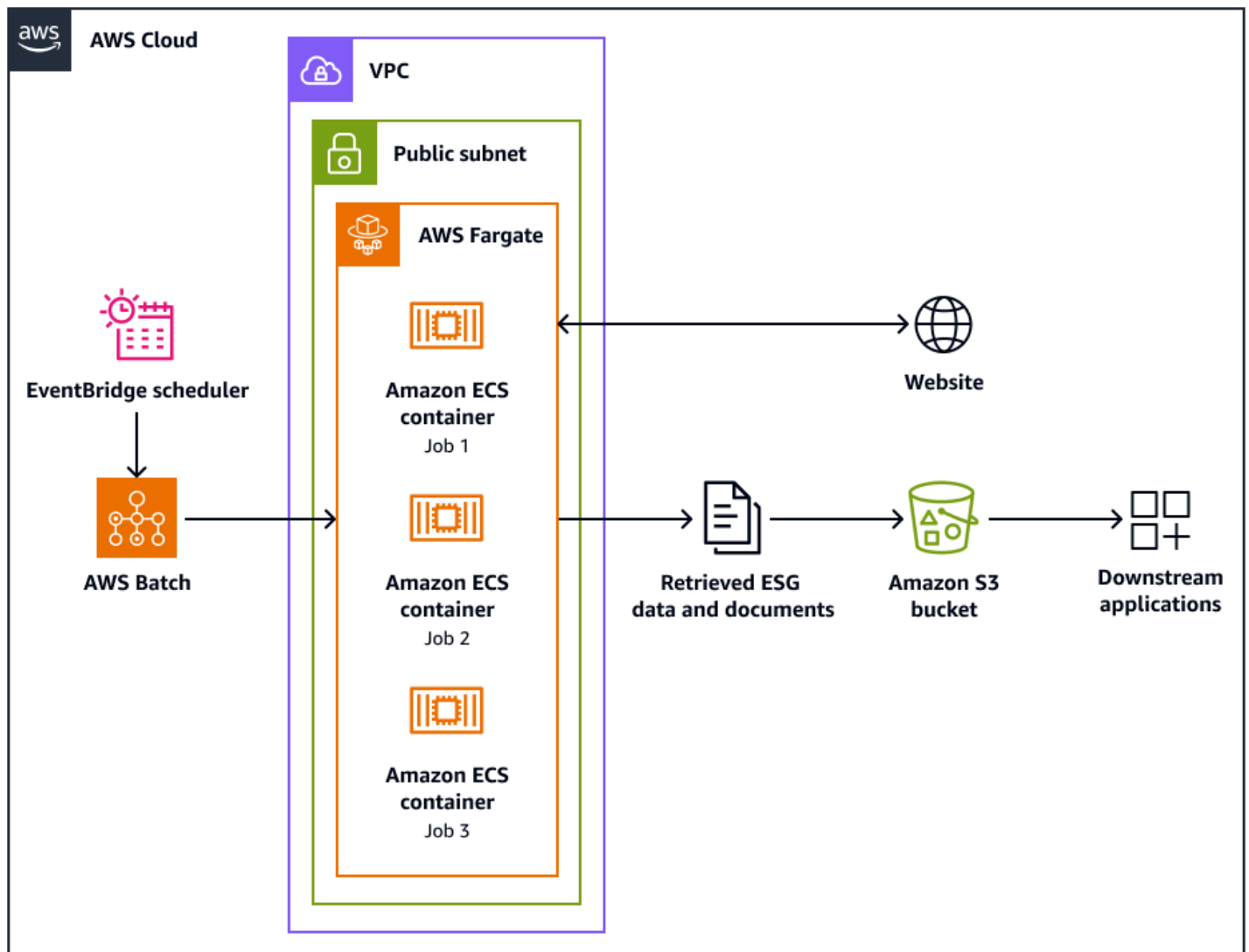
## Targeted business outcomes

The following are common reasons that companies use ESG data:

- **Risk management** – ESG data helps you identify and mitigate potential risks related to environmental, social, and governance issues.
- **Investor attraction** – Many investors now consider ESG factors when making investment decisions. They view strong ESG practices as indicators of long-term sustainability and profitability.
- **Reputation management** – Good ESG performance can enhance a company's reputation among customers, employees, and the general public.
- **Regulatory compliance** – As ESG-related regulations increase, adopting ESG practices helps companies stay ahead of compliance requirements.
- **Innovation and efficiency** – Focusing on ESG factors can drive innovation in products, services, and operations. This leads to improved efficiency and cost savings.
- **Competitive advantage** – Strong ESG performance can differentiate a company from its competitors and open up new market opportunities.
- **Stakeholder engagement** – ESG practices help companies better engage with and meet the expectations of various stakeholders, including employees, customers, and local communities.

# Architecture for a scalable web crawling system on AWS

The following architecture diagram shows a web crawler system that is designed to ethically extract environmental, social, and governance (ESG) data from websites. You use a Python-based crawler that is optimized for AWS infrastructure. You use AWS Batch to orchestrate the large-scale crawling jobs and use Amazon Simple Storage Service (Amazon S3) for storage. Downstream applications can ingest and store the data from the Amazon S3 bucket.



The diagram shows the following workflow:

1. Amazon EventBridge Scheduler initiates the crawling process at an interval that you schedule.

2. AWS Batch manages the execution of the web crawler jobs. The AWS Batch job queue holds and orchestrates the pending crawling jobs.
3. The web crawling jobs run in Amazon Elastic Container Service (Amazon ECS) containers on AWS Fargate. The jobs run in a public subnet of a virtual private cloud (VPC).
4. The web crawler crawls the target website and retrieves ESG data and documents, such as PDF, CSV, or other document files.
5. The web crawler stores the retrieved data and raw files in an Amazon S3 bucket.
6. Other systems or applications ingest or process the stored data and files in the Amazon S3 bucket.

## Web crawler design and operations

Some websites are specifically designed to be run on desktops or on mobile devices. The web crawler is designed to support the use of a desktop user agent or a mobile user agent. These agents help you successfully make requests to the target website.

After the web crawler is initialized, it performs the following operations:

1. The web crawler calls the `setup()` method. This method fetches and parses the `robots.txt` file.

### Note

You can also configure the web crawler to fetch and parse the sitemap.

2. The web crawler processes the `robots.txt` file. If a crawl delay is specified in the `robots.txt` file, the web crawler extracts the crawl delay for the desktop user agent. If a crawl delay is not specified in the `robots.txt` file, then the web crawler uses a random delay.
3. The web crawler calls the `crawl()` method, which initiates the crawling process. If no URLs are in the queue, it adds the start URL.

### Note

The crawler continues until it reaches the maximum number of pages or runs out of URLs to crawl.

4. The crawler processes the URLs. For each URL in the queue, the crawler checks if the URL has already been crawled.



5. If a URL hasn't been crawled, the crawler calls the `crawl_url()` method as follows:
  - a. The crawler checks the `robots.txt` file to determine whether it can use the desktop user agent to crawl the URL.
  - b. If allowed, the crawler attempts to crawl the URL by using the desktop user agent.
  - c. If not allowed or if the desktop user agent fails to crawl, then the crawler checks the `robots.txt` file to determine whether it can use the mobile user agent to crawl the URL.
  - d. If allowed, the crawler attempts to crawl the URL by using the mobile user agent.
6. The crawler calls the `attempt_crawl()` method, which retrieves and processes the content. The crawler sends a GET request to the URL with appropriate headers. If the request fails, the crawler uses retry logic.
7. If the file is in HTML format, the crawler calls the `extract_esg_data()` method. It uses [Beautiful Soup](#) to parse the HTML content. It extracts environmental, social, and governance (ESG) data by using keyword matching.

If the file is a PDF, the crawler calls the `save_pdf()` method. The crawler downloads and saves the PDF file to the Amazon S3 bucket.
8. The crawler calls the `extract_news_links()` method. This finds and stores links to news articles, press releases, and blog posts.
9. The crawler calls the `extract_pdf_links()` method. This identifies and stores links to PDF documents.
10. The crawler calls the `is_relevant_to_sustainable_finance()` method. This checks whether the news or articles are related to sustainable finance by using predefined keywords.
11. After each crawl attempt, the crawler implements a delay by using the `delay()` method. If a delay was specified in the `robots.txt` file, it uses that value. Otherwise, it uses a random delay between 1 and 3 seconds.
12. The crawler calls the `save_esg_data()` method to save the ESG data to a CSV file. The CSV file is saved in the Amazon S3 bucket.
13. The crawler calls the `save_news_links()` method to save the news links to a CSV file, including relevance information. The CSV file is saved in the Amazon S3 bucket.
14. The crawler calls the `save_pdf_links()` method to save the PDF links to a CSV file. The CSV file is saved in the Amazon S3 bucket.

## Batching and data processing

The crawling process is organized and performed in a structured manner. AWS Batch assigns the jobs for each company so that they run in parallel, in batches. Each batch focuses on a single company's domain and subdomains, as you've identified them in your dataset. However, jobs in the same batch run sequentially so that they do not inundate the website with too many requests. This helps the application to manage the crawling workload more efficiently and make sure that all relevant data is captured for each company.

By organizing the web crawling into company-specific batches, this containerizes the collected data. This helps prevent the data from one company from being mixed with data from other companies.

Batching helps the application efficiently gather data from the web, while maintaining a clear structure and separation of information based on the target companies and their respective web domains. This approach helps to ensure the integrity and usability of the data collected, as it is neatly organized and associated with the appropriate company and domains.

# Building a scalable web crawling system on AWS

This section describes how to build the web crawler described in the [Architecture](#) section. It includes a systematic approach to creating a robust dataset of companies and their associated web properties. This dataset serves as the foundation for your crawling activities. Then, this section describes how to build an ethical web crawler in Python.

## Topics

- [Preparing a dataset](#)
- [Building the web crawler](#)
- [Building the AWS infrastructure](#)

## Preparing a dataset

If you have not done so already, prepare a detailed dataset of the websites that you want to collect information from. This dataset should include website URL domain names and relevant subdomain names. This section provides a step-by-step process for building this dataset.

### To prepare a dataset

1. **Define the scope** – Determine the industry or sectors you're focusing on. Decide how many companies to include. And define any criteria that you want to collect about these companies, such as the number of employees, location, or revenue.
2. **Identify data sources** – Identify what sources of information you can use to collect information about these companies. Examples include business directories (such as [Crunchbase](#), [Bloomberg](#), or [Forbes](#)), stock exchanges (such as NYSE and NASDAQ), industry-specific associations or publications, or government databases (such as SEC filings).
3. **Create a table** – In your preferred tool, such as Microsoft Excel, Google Sheets, or a database management system, create a table for collecting criteria about each company. Include a column for each criterion. At a minimum, include columns for the company name, primary domain, subdomains, industry, size, and location.
4. **Collect initial company information** – Collect the following information about each company and enter it into the table that you created:
  - Company name

- Industry or sector
  - Company size (number of employees)
  - Revenue
  - Location of company headquarters
5. **Gather domain information** – For each company, extract the primary domain name from the main website URL, such as `example.com`. You can verify the domain information by using a WHOIS domain lookup tool.
  6. **Gather subdomain information** – For each company, research the registered subdomains, such as `blog.example.com`. You can use subdomain enumeration tools, such as [Sublist3r](#), [OWASP Amass](#), or [Subfinder](#). You can perform Google dorking (by searching `site:example.com`), check DNS records by using a `dig` command or a DNS lookup tool, or you can analyze SSL or TLS certificates.
  7. **Validate and clean the data** — Review, verify, and standardize the data that you've collected. For example, remove any duplicate entries, remove unnecessary URL information from domains and subdomains, and verify that all domains and subdomains are active.
  8. **(Optional) Categorize the subdomains** – You can categorize the subdomains into types. The following are some examples of categories you might encounter:
    - Blogs, such as `blog.example.com`
    - Support or help, such as `support.example.com` or `help.example.com`
    - E-commerce, such as `shop.example.com` or `store.example.com`
    - Developer resources, such as `dev.example.com` or `api.example.com`
    - Regions or locations, such as `us.example.com` or `uk.example.com`
  9. **(Optional) Add relevant metadata** – You can record any relevant metadata in the dataset. For example, you can add the last updated date, the source of information, or your confidence score for subdomain accuracy.
  10. **Implement version control** – Use a version control system, such as Git, to track changes to the table over time. Back up the dataset regularly.
  11. **Maintain the table** – Set up a schedule, such as quarterly, for updating the table. Standardize and implement a process for adding new companies or removing those you no longer need. When possible, automate discovery of subdomains.

# Building the web crawler

As described in the [Architecture](#) section, the application runs in batches—one for each company.

## Topics

- [Capturing and processing the robots.txt file](#)
- [Capturing and processing the sitemap](#)
- [Designing the crawler](#)

## Capturing and processing the robots.txt file

After you prepare the dataset, you need to confirm whether the domain has a [robots.txt](#) file. For web crawlers and other bots, the robots.txt file indicates which sections of the website they are allowed to visit. Honoring the instructions in this file is an important best practice for ethically crawling websites. For more information, see [Best practices for ethical web crawlers](#) in this guide.

### To capture and process the robots.txt file

1. If you haven't done so already, install the `requests` library by running the following command in a terminal:

```
pip install requests
```

2. Run the following script. This script does the following:
  - It defines a `check_robots_txt` function that takes a domain as input.
  - It constructs the full URL for the robots.txt file.
  - It sends a GET request to the URL for the robots.txt file.
  - If the request is successful (status code 200), then a robots.txt file exists.
  - If the request fails or returns a different status code, then a robots.txt file doesn't exist or isn't accessible.

```
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**

This script handles exceptions for network errors or other issues.

### 3. If a robots.txt file exists, use the following script to download it:

```
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**

These scripts can be customized or modified according to your use case. You can also combine these scripts.

## Capturing and processing the sitemap

Next, you need to process the [sitemap](#). You can use the sitemap to focus crawling on important pages. This improves the crawling efficiency. For more information, see [Best practices for ethical web crawlers](#) in this guide.

### To capture and process the sitemap

- Run the following script. This script defines a `check_and_download_sitemap` function that:
  - Accepts a base URL, an optional sitemap URL from robots.txt, and a user-agent string.
  - Checks multiple potential sitemap locations, including the one from robots.txt (if provided).
  - Attempts to download the sitemap from each location.
  - Verifies that the downloaded content is in XML format.
  - Calls the `parse_sitemap` function to extract the URLs. This function:
    - Parses the XML content of the sitemap.
    - Handles both regular sitemaps and sitemap index files.
    - Recursively fetches sub-sitemaps if a sitemap index is encountered.

```
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("...")
```

## Designing the crawler

Next, you design the web crawler. The crawler is designed to follow the best practices described in [Best practices for ethical web crawlers](#) in this guide. This EthicalCrawler class demonstrates several key principles of ethical crawling:

- **Fetching and parsing the robots.txt file** – The crawler fetches the robots.txt file for the target website.
- **Respecting crawling permissions** – Before crawling any URL, the crawler checks if the rules in the robots.txt file allow crawling for that URL. If a URL is disallowed, then the crawler skips it and moves to the next URL.
- **Honoring crawl delay** – The crawler checks for a crawl-delay directive in the robots.txt file. If one is specified, the crawler uses this delay between requests. Otherwise, it uses a default delay.
- **User-agent identification** – The crawler uses a custom user-agent string to identify itself to websites. If needed, website owners can set specific rules to restrict or allow your crawler.
- **Error handling and graceful degradation** – If the robots.txt file can't be fetched or parsed, the crawler proceeds with conservative default rules. It handles network errors and non-200 HTTP responses.
- **Limited crawling** – To avoid overwhelming the server, there is a limit to how many pages can be crawled.

The following script is pseudocode that explains how the web crawler works:

```
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.")
```

## To build an advanced, ethical web crawler that collects ESG data

1. Copy the following code sample for the advanced ethical web crawler used in this system:

```
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. Set up the various attributes, including user agents, empty collections for URLs, and data storage lists.
3. Adjust the keywords and relevance criteria in the `is_relevant_to_sustainable_finance()` method to match your specific needs.
4. Make sure that the robots.txt file permits crawling the website and that you are using the crawl delay and user agent specified in the robots.txt file.

5. Consider making the following customizations to the provided web crawler script, as needed for your organization:
  - Implement the `fetch_sitemap()` method for more efficient URL discovery.
  - Enhance error logging and handling for production use.
  - Implement more sophisticated content relevance analysis.
  - Add depth and breadth controls to limit crawling scope.

## Building the AWS infrastructure

There are many AWS services that you can use to build the web crawling infrastructure. The [Architecture](#) section of this guide includes one proposed solution. We recommend that you consider using the following AWS services to build the supporting infrastructure for your web crawler:

- Use Amazon Virtual Private Cloud (Amazon VPC) to create the [VPC](#) and [subnets](#).
- Initiate the crawling process by using [Amazon EventBridge Scheduler](#).
- Manage the web crawler jobs by using AWS Batch [jobs](#) and [job queues](#).
- Use one of the following solutions to run the web crawler jobs:
  - Amazon Elastic Container Service (Amazon ECS) containers on [AWS Fargate](#)
  - Amazon Elastic Compute Cloud (Amazon EC2) [instances](#)

### Note

If your application can handle disruptions, consider using Amazon EC2 [Spot Instances](#) through [Spot Fleet](#). Fleets of Spot Instances can help you save significantly on compute costs.

- AWS Lambda [functions](#)
- Store the retrieved data and raw files in an Amazon Simple Storage Service (Amazon S3) [bucket](#).



# Best practices for ethical web crawlers

This section discusses best practices and key ethical considerations for building a web-crawling application that collects environmental, social, and governance (ESG) data. By adhering to these best practices, you can protect your project and organization and contribute to a more responsible and sustainable web ecosystem. This approach helps you access valuable data and use it for research, business, and innovation in a way that respects all stakeholders.

## Robots.txt compliance

The robots.txt file is used on websites to communicate with web crawlers and bots about which parts of the website should or should not be accessed or crawled. When a web crawler encounters a robots.txt file on a website, it parses the instructions and adjusts its crawling behavior accordingly. This prevents the crawler from violating the website owner's instructions and maintains a cooperative relationship between the website and the crawler. Therefore, the robots.txt file helps with access control, the protection of sensitive content, load management, and legal compliance.

We recommend that you adhere to the following best practices:

- Always check and respect the rules in the robots.txt file.
- Before crawling any URL, check the rules for both desktop and mobile user agents.
- If the website allows only mobile user agents, use a different agent header, such as a mobile agent header, for your request.

The absence of a robots.txt file doesn't necessarily mean you can't or shouldn't crawl a website. Crawling should always be done responsibly, respecting the website's resources and the owner's implicit rights. The following are recommended best practices when a robots.txt is not present:

- Assume crawling is allowed, but proceed with caution.
- Implement polite crawling practices.
- Consider reaching out to the website owner for permission if you plan on performing extensive crawling.

## Crawl-rate limiting

Use a reasonable crawl rate to avoid overwhelming the server. Implement delays between requests, either as specified by the robots.txt file or by using a random delay. For small or medium-sized websites, 1 request every 10–15 seconds might be appropriate. For larger websites or those with explicit crawl permissions, 1–2 requests per second might be appropriate.

## User-agent transparency

Identify your crawler in the user-agent header. This HTTP header information is intended to identify the device that is requesting the content. Commonly, the word *bot* is included in the name of the agent. Crawlers and other bots sometimes use an important field in the header to include contact information.

## Efficient crawling

Use the sitemap, which is developed by the website owner, in order to focus on important pages.

## Adaptive approach

Program the crawler to switch to a mobile user agent if the desktop version is unsuccessful. This can provide the crawler access and reduce the strain on the website's server.

## Error handling

Make sure that the crawler handles various HTTP status codes appropriately. For example, the crawler should pause if it encounters a 429 status code ("Too many requests"). If the crawler continuously receives 403 status codes ("Forbidden"), then consider stopping crawling.

## Crawling in batches

We recommend that you do the following:

- Instead of crawling all the URLs at once, divide the task into smaller batches. This can help distribute the load and reduce the risk of encountering issues, such as timeouts or resource constraints.

- If the overall crawling task is expected to be long-running, consider dividing it into multiple smaller, more manageable tasks. This can make the process more scalable and resilient.
- If the number of URLs to crawl is relatively small, consider using a serverless solution, such as AWS Lambda. Lambda functions can be a good fit for short-lived, event-driven tasks because they automatically scale and handle resource management.

## Security

For web-crawling compute tasks, we recommend that you configure the environment to allow outbound traffic only. This helps enhance security by minimizing the attack surface and reducing the risk of unauthorized inbound access. Allowing only outbound connections permits the crawling process to communicate with the target websites and retrieve the necessary data, and it restricts any inbound traffic that could potentially compromise the system.

## Other considerations

Review the following additional considerations and best practices:

- Check for crawling guidelines in the website's terms of service or privacy policy.
- Look for meta tags in the HTML that might provide crawling directives.
- Be aware of legal restrictions in your jurisdiction regarding data collection and use.
- Be prepared to stop crawling if requested by the website owner.

# FAQ

## What if a robots.txt file is not available?

The absence of a robots.txt file doesn't necessarily mean you can't or shouldn't crawl a website. Crawling should always be done responsibly, respecting the website's resources and the implicit rights of the website owner.

## What if a sitemaps.xml file is not available?

Depending on the requirement, you can do one of the following:

- **Search for HTML sitemaps** – Look for an HTML sitemap page that lists the important pages on the website. These are often linked in the footer.
- **Crawl from the homepage** – Start crawling from the homepage and follow internal links to discover other pages.
- **Analyze URL patterns** – Analyze the website's URL structure to identify patterns and programmatically generate potential URLs.
- **Review the robots.txt file** – Check the robots.txt file for any disallowed pages or directories. These can provide clues about the site structure.
- **Review API endpoints** – Some websites offer API endpoints that can be used to retrieve content and structure information.
- **Check search engine results** – Use search engines to find indexed pages of the website by using the [site: search operator](#), such as `site:example.com`.
- **Analyze backlinks** – Analyze backlinks to the website to discover important pages that other sites are linking to.
- **Review web archives** – Check internet archives, such as the [Wayback Machine](#), for older versions of the site that might have had sitemaps or different structures.
- **Look for content management system (CMS) patterns** – If you can identify the CMS, use common URL patterns associated with that system.
- **Confirm JavaScript rendering** – If the site heavily relies on JavaScript, make sure that your crawler can render JavaScript to discover dynamically loaded content. For some websites, the sitemap.xml file loads after JavaScript rendering is enabled.

## Can I use a serverless solution instead of Amazon EC2 or Amazon ECS?

Yes. [AWS Lambda](#) functions for web crawling can be a viable option, especially for smaller-scale or more modular crawling tasks. However, for large-scale, long-running crawling operations, a more traditional approach that uses Amazon Elastic Compute Cloud (Amazon EC2) instances or Amazon Elastic Container Service (Amazon ECS) might be more suitable. It's important to carefully evaluate your specific requirements and trade-offs when choosing the right compute service for your web crawling needs.

## Why is the crawler getting a 403 status code?

HTTP 403 is an HTTP status code that means access to the requested resource is forbidden. If the request was correct, then the server understood the request and will not fulfill it. To prevent a 403 status code, you can do the following:

- Limit your crawl rate.
- Check whether the sitemap or robots.txt file allows the crawler to access the URL.
- Try with a mobile user agent instead of a desktop user agent.

If none of the above work, you should respect the decision of the website owners and not crawl the page.

## Next steps and resources

After collecting the raw environmental, social, and governance (ESG) data, you can do the following to extract meaningful information from the data:

1. **Clean the data** – The data might include a significant amount of irrelevant information that is unrelated to ESG factors and financial data. It is important to remove this irrelevant data and retain only the information needed to perform the required analytics. You can use tools, such as like [yfinance](#), to help clean the data.
2. **Extract and transform the data** – Extract the relevant features or variables from the raw data and transform them into a format that is suitable for analysis. You can transform the data into tabular format for better readability and clarity. You can use a library, such as [pandas](#), to refine the data. You can also use feature engineering, data normalization, and derived metrics to transform the data.
3. **Perform analytics** – You can perform various analytical tasks. This might include generating descriptive statistics, creating data visualizations, and conducting exploratory data analysis to gain insights into the ESG performance of the companies.
4. **Apply machine learning** – You can use the cleaned and transformed data to train machine learning models. These models can help you identify companies that are currently demonstrating financial sustainability and project their future sustainability performance.

By using the web crawler and this data evaluation process, you can effectively gain a comprehensive understanding of the sustainability practices and financial performance of the companies you are evaluating. You can use this information to inform investment decisions, track progress, and support sustainable business practices.

## Resources

- [What is a web crawler?](#) (Cloudflare website)
- [Guide to ESG investment](#) (Investopedia website)

## Tools

- [Beautiful Soup](#) (Beautiful Soup documentation)

- [Handling tabular, relational data](#) (pandas website)

## Document history

The following table describes significant changes to this guide. If you want to be notified about future updates, you can subscribe to an [RSS feed](#).

| Change                              | Description | Date            |
|-------------------------------------|-------------|-----------------|
| <a href="#">Initial publication</a> | —           | January 6, 2025 |



# AWS Prescriptive Guidance glossary

The following are commonly used terms in strategies, guides, and patterns provided by AWS Prescriptive Guidance. To suggest entries, please use the **Provide feedback** link at the end of the glossary.

## Numbers

### 7 Rs

Seven common migration strategies for moving applications to the cloud. These strategies build upon the 5 Rs that Gartner identified in 2011 and consist of the following:

- Refactor/re-architect – Move an application and modify its architecture by taking full advantage of cloud-native features to improve agility, performance, and scalability. This typically involves porting the operating system and database. Example: Migrate your on-premises Oracle database to the Amazon Aurora PostgreSQL-Compatible Edition.
- Replatform (lift and reshape) – Move an application to the cloud, and introduce some level of optimization to take advantage of cloud capabilities. Example: Migrate your on-premises Oracle database to Amazon Relational Database Service (Amazon RDS) for Oracle in the AWS Cloud.
- Repurchase (drop and shop) – Switch to a different product, typically by moving from a traditional license to a SaaS model. Example: Migrate your customer relationship management (CRM) system to Salesforce.com.
- Rehost (lift and shift) – Move an application to the cloud without making any changes to take advantage of cloud capabilities. Example: Migrate your on-premises Oracle database to Oracle on an EC2 instance in the AWS Cloud.
- Relocate (hypervisor-level lift and shift) – Move infrastructure to the cloud without purchasing new hardware, rewriting applications, or modifying your existing operations. You migrate servers from an on-premises platform to a cloud service for the same platform. Example: Migrate a Microsoft Hyper-V application to AWS.
- Retain (revisit) – Keep applications in your source environment. These might include applications that require major refactoring, and you want to postpone that work until a later time, and legacy applications that you want to retain, because there's no business justification for migrating them.

- Retire – Decommission or remove applications that are no longer needed in your source environment.

## A

### ABAC

See [attribute-based access control](#).

### abstracted services

See [managed services](#).

### ACID

See [atomicity, consistency, isolation, durability](#).

### active-active migration

A database migration method in which the source and target databases are kept in sync (by using a bidirectional replication tool or dual write operations), and both databases handle transactions from connecting applications during migration. This method supports migration in small, controlled batches instead of requiring a one-time cutover. It's more flexible but requires more work than [active-passive migration](#).

### active-passive migration

A database migration method in which the source and target databases are kept in sync, but only the source database handles transactions from connecting applications while data is replicated to the target database. The target database doesn't accept any transactions during migration.

### aggregate function

A SQL function that operates on a group of rows and calculates a single return value for the group. Examples of aggregate functions include SUM and MAX.

### AI

See [artificial intelligence](#).

### AIOps

See [artificial intelligence operations](#).

## anonymization

The process of permanently deleting personal information in a dataset. Anonymization can help protect personal privacy. Anonymized data is no longer considered to be personal data.

## anti-pattern

A frequently used solution for a recurring issue where the solution is counter-productive, ineffective, or less effective than an alternative.

## application control

A security approach that allows the use of only approved applications in order to help protect a system from malware.

## application portfolio

A collection of detailed information about each application used by an organization, including the cost to build and maintain the application, and its business value. This information is key to [the portfolio discovery and analysis process](#) and helps identify and prioritize the applications to be migrated, modernized, and optimized.

## artificial intelligence (AI)

The field of computer science that is dedicated to using computing technologies to perform cognitive functions that are typically associated with humans, such as learning, solving problems, and recognizing patterns. For more information, see [What is Artificial Intelligence?](#)

## artificial intelligence operations (AIOps)

The process of using machine learning techniques to solve operational problems, reduce operational incidents and human intervention, and increase service quality. For more information about how AIOps is used in the AWS migration strategy, see the [operations integration guide](#).

## asymmetric encryption

An encryption algorithm that uses a pair of keys, a public key for encryption and a private key for decryption. You can share the public key because it isn't used for decryption, but access to the private key should be highly restricted.

## atomicity, consistency, isolation, durability (ACID)

A set of software properties that guarantee the data validity and operational reliability of a database, even in the case of errors, power failures, or other problems.

## attribute-based access control (ABAC)

The practice of creating fine-grained permissions based on user attributes, such as department, job role, and team name. For more information, see [ABAC for AWS](#) in the AWS Identity and Access Management (IAM) documentation.

## authoritative data source

A location where you store the primary version of data, which is considered to be the most reliable source of information. You can copy data from the authoritative data source to other locations for the purposes of processing or modifying the data, such as anonymizing, redacting, or pseudonymizing it.

## Availability Zone

A distinct location within an AWS Region that is insulated from failures in other Availability Zones and provides inexpensive, low-latency network connectivity to other Availability Zones in the same Region.

## AWS Cloud Adoption Framework (AWS CAF)

A framework of guidelines and best practices from AWS to help organizations develop an efficient and effective plan to move successfully to the cloud. AWS CAF organizes guidance into six focus areas called perspectives: business, people, governance, platform, security, and operations. The business, people, and governance perspectives focus on business skills and processes; the platform, security, and operations perspectives focus on technical skills and processes. For example, the people perspective targets stakeholders who handle human resources (HR), staffing functions, and people management. For this perspective, AWS CAF provides guidance for people development, training, and communications to help ready the organization for successful cloud adoption. For more information, see the [AWS CAF website](#) and the [AWS CAF whitepaper](#).

## AWS Workload Qualification Framework (AWS WQF)

A tool that evaluates database migration workloads, recommends migration strategies, and provides work estimates. AWS WQF is included with AWS Schema Conversion Tool (AWS SCT). It analyzes database schemas and code objects, application code, dependencies, and performance characteristics, and provides assessment reports.

## B

### bad bot

A [bot](#) that is intended to disrupt or cause harm to individuals or organizations.

### BCP

See [business continuity planning](#).

### behavior graph

A unified, interactive view of resource behavior and interactions over time. You can use a behavior graph with Amazon Detective to examine failed logon attempts, suspicious API calls, and similar actions. For more information, see [Data in a behavior graph](#) in the Detective documentation.

### big-endian system

A system that stores the most significant byte first. See also [endianness](#).

### binary classification

A process that predicts a binary outcome (one of two possible classes). For example, your ML model might need to predict problems such as "Is this email spam or not spam?" or "Is this product a book or a car?"

### bloom filter

A probabilistic, memory-efficient data structure that is used to test whether an element is a member of a set.

### blue/green deployment

A deployment strategy where you create two separate but identical environments. You run the current application version in one environment (blue) and the new application version in the other environment (green). This strategy helps you quickly roll back with minimal impact.

### bot

A software application that runs automated tasks over the internet and simulates human activity or interaction. Some bots are useful or beneficial, such as web crawlers that index information on the internet. Some other bots, known as *bad bots*, are intended to disrupt or cause harm to individuals or organizations.

## botnet

Networks of [bots](#) that are infected by [malware](#) and are under the control of a single party, known as a *bot herder* or *bot operator*. Botnets are the best-known mechanism to scale bots and their impact.

## branch

A contained area of a code repository. The first branch created in a repository is the *main branch*. You can create a new branch from an existing branch, and you can then develop features or fix bugs in the new branch. A branch you create to build a feature is commonly referred to as a *feature branch*. When the feature is ready for release, you merge the feature branch back into the main branch. For more information, see [About branches](#) (GitHub documentation).

## break-glass access

In exceptional circumstances and through an approved process, a quick means for a user to gain access to an AWS account that they don't typically have permissions to access. For more information, see the [Implement break-glass procedures](#) indicator in the AWS Well-Architected guidance.

## brownfield strategy

The existing infrastructure in your environment. When adopting a brownfield strategy for a system architecture, you design the architecture around the constraints of the current systems and infrastructure. If you are expanding the existing infrastructure, you might blend brownfield and [greenfield](#) strategies.

## buffer cache

The memory area where the most frequently accessed data is stored.

## business capability

What a business does to generate value (for example, sales, customer service, or marketing). Microservices architectures and development decisions can be driven by business capabilities. For more information, see the [Organized around business capabilities](#) section of the [Running containerized microservices on AWS](#) whitepaper.

## business continuity planning (BCP)

A plan that addresses the potential impact of a disruptive event, such as a large-scale migration, on operations and enables a business to resume operations quickly.

# C

## CAF

See [AWS Cloud Adoption Framework](#).

## canary deployment

The slow and incremental release of a version to end users. When you are confident, you deploy the new version and replace the current version in its entirety.

## CCoE

See [Cloud Center of Excellence](#).

## CDC

See [change data capture](#).

## change data capture (CDC)

The process of tracking changes to a data source, such as a database table, and recording metadata about the change. You can use CDC for various purposes, such as auditing or replicating changes in a target system to maintain synchronization.

## chaos engineering

Intentionally introducing failures or disruptive events to test a system's resilience. You can use [AWS Fault Injection Service \(AWS FIS\)](#) to perform experiments that stress your AWS workloads and evaluate their response.

## CI/CD

See [continuous integration and continuous delivery](#).

## classification

A categorization process that helps generate predictions. ML models for classification problems predict a discrete value. Discrete values are always distinct from one another. For example, a model might need to evaluate whether or not there is a car in an image.

## client-side encryption

Encryption of data locally, before the target AWS service receives it.

## Cloud Center of Excellence (CCoE)

A multi-disciplinary team that drives cloud adoption efforts across an organization, including developing cloud best practices, mobilizing resources, establishing migration timelines, and leading the organization through large-scale transformations. For more information, see the [CCoE posts](#) on the AWS Cloud Enterprise Strategy Blog.

## cloud computing

The cloud technology that is typically used for remote data storage and IoT device management. Cloud computing is commonly connected to [edge computing](#) technology.

## cloud operating model

In an IT organization, the operating model that is used to build, mature, and optimize one or more cloud environments. For more information, see [Building your Cloud Operating Model](#).

## cloud stages of adoption

The four phases that organizations typically go through when they migrate to the AWS Cloud:

- Project – Running a few cloud-related projects for proof of concept and learning purposes
- Foundation – Making foundational investments to scale your cloud adoption (e.g., creating a landing zone, defining a CCoE, establishing an operations model)
- Migration – Migrating individual applications
- Re-invention – Optimizing products and services, and innovating in the cloud

These stages were defined by Stephen Orban in the blog post [The Journey Toward Cloud-First & the Stages of Adoption](#) on the AWS Cloud Enterprise Strategy blog. For information about how they relate to the AWS migration strategy, see the [migration readiness guide](#).

## CMDB

See [configuration management database](#).

## code repository

A location where source code and other assets, such as documentation, samples, and scripts, are stored and updated through version control processes. Common cloud repositories include GitHub or Bitbucket Cloud. Each version of the code is called a *branch*. In a microservice structure, each repository is devoted to a single piece of functionality. A single CI/CD pipeline can use multiple repositories.



## cold cache

A buffer cache that is empty, not well populated, or contains stale or irrelevant data. This affects performance because the database instance must read from the main memory or disk, which is slower than reading from the buffer cache.

## cold data

Data that is rarely accessed and is typically historical. When querying this kind of data, slow queries are typically acceptable. Moving this data to lower-performing and less expensive storage tiers or classes can reduce costs.

## computer vision (CV)

A field of [AI](#) that uses machine learning to analyze and extract information from visual formats such as digital images and videos. For example, Amazon SageMaker AI provides image processing algorithms for CV.

## configuration drift

For a workload, a configuration change from the expected state. It might cause the workload to become noncompliant, and it's typically gradual and unintentional.

## configuration management database (CMDB)

A repository that stores and manages information about a database and its IT environment, including both hardware and software components and their configurations. You typically use data from a CMDB in the portfolio discovery and analysis stage of migration.

## conformance pack

A collection of AWS Config rules and remediation actions that you can assemble to customize your compliance and security checks. You can deploy a conformance pack as a single entity in an AWS account and Region, or across an organization, by using a YAML template. For more information, see [Conformance packs](#) in the AWS Config documentation.

## continuous integration and continuous delivery (CI/CD)

The process of automating the source, build, test, staging, and production stages of the software release process. CI/CD is commonly described as a pipeline. CI/CD can help you automate processes, improve productivity, improve code quality, and deliver faster. For more information, see [Benefits of continuous delivery](#). CD can also stand for *continuous deployment*. For more information, see [Continuous Delivery vs. Continuous Deployment](#).

## CV

See [computer vision](#).

## D

### data at rest

Data that is stationary in your network, such as data that is in storage.

### data classification

A process for identifying and categorizing the data in your network based on its criticality and sensitivity. It is a critical component of any cybersecurity risk management strategy because it helps you determine the appropriate protection and retention controls for the data. Data classification is a component of the security pillar in the AWS Well-Architected Framework. For more information, see [Data classification](#).

### data drift

A meaningful variation between the production data and the data that was used to train an ML model, or a meaningful change in the input data over time. Data drift can reduce the overall quality, accuracy, and fairness in ML model predictions.

### data in transit

Data that is actively moving through your network, such as between network resources.

### data mesh

An architectural framework that provides distributed, decentralized data ownership with centralized management and governance.

### data minimization

The principle of collecting and processing only the data that is strictly necessary. Practicing data minimization in the AWS Cloud can reduce privacy risks, costs, and your analytics carbon footprint.

### data perimeter

A set of preventive guardrails in your AWS environment that help make sure that only trusted identities are accessing trusted resources from expected networks. For more information, see [Building a data perimeter on AWS](#).

## data preprocessing

To transform raw data into a format that is easily parsed by your ML model. Preprocessing data can mean removing certain columns or rows and addressing missing, inconsistent, or duplicate values.

## data provenance

The process of tracking the origin and history of data throughout its lifecycle, such as how the data was generated, transmitted, and stored.

## data subject

An individual whose data is being collected and processed.

## data warehouse

A data management system that supports business intelligence, such as analytics. Data warehouses commonly contain large amounts of historical data, and they are typically used for queries and analysis.

## database definition language (DDL)

Statements or commands for creating or modifying the structure of tables and objects in a database.

## database manipulation language (DML)

Statements or commands for modifying (inserting, updating, and deleting) information in a database.

## DDL

See [database definition language](#).

## deep ensemble

To combine multiple deep learning models for prediction. You can use deep ensembles to obtain a more accurate prediction or for estimating uncertainty in predictions.

## deep learning

An ML subfield that uses multiple layers of artificial neural networks to identify mapping between input data and target variables of interest.

## defense-in-depth

An information security approach in which a series of security mechanisms and controls are thoughtfully layered throughout a computer network to protect the confidentiality, integrity, and availability of the network and the data within. When you adopt this strategy on AWS, you add multiple controls at different layers of the AWS Organizations structure to help secure resources. For example, a defense-in-depth approach might combine multi-factor authentication, network segmentation, and encryption.

## delegated administrator

In AWS Organizations, a compatible service can register an AWS member account to administer the organization's accounts and manage permissions for that service. This account is called the *delegated administrator* for that service. For more information and a list of compatible services, see [Services that work with AWS Organizations](#) in the AWS Organizations documentation.

## deployment

The process of making an application, new features, or code fixes available in the target environment. Deployment involves implementing changes in a code base and then building and running that code base in the application's environments.

## development environment

See [environment](#).

## detective control

A security control that is designed to detect, log, and alert after an event has occurred. These controls are a second line of defense, alerting you to security events that bypassed the preventative controls in place. For more information, see [Detective controls](#) in *Implementing security controls on AWS*.

## development value stream mapping (DVSM)

A process used to identify and prioritize constraints that adversely affect speed and quality in a software development lifecycle. DVSM extends the value stream mapping process originally designed for lean manufacturing practices. It focuses on the steps and teams required to create and move value through the software development process.

## digital twin

A virtual representation of a real-world system, such as a building, factory, industrial equipment, or production line. Digital twins support predictive maintenance, remote monitoring, and production optimization.

## dimension table

In a [star schema](#), a smaller table that contains data attributes about quantitative data in a fact table. Dimension table attributes are typically text fields or discrete numbers that behave like text. These attributes are commonly used for query constraining, filtering, and result set labeling.

## disaster

An event that prevents a workload or system from fulfilling its business objectives in its primary deployed location. These events can be natural disasters, technical failures, or the result of human actions, such as unintentional misconfiguration or a malware attack.

## disaster recovery (DR)

The strategy and process you use to minimize downtime and data loss caused by a [disaster](#). For more information, see [Disaster Recovery of Workloads on AWS: Recovery in the Cloud](#) in the AWS Well-Architected Framework.

## DML

See [database manipulation language](#).

## domain-driven design

An approach to developing a complex software system by connecting its components to evolving domains, or core business goals, that each component serves. This concept was introduced by Eric Evans in his book, *Domain-Driven Design: Tackling Complexity in the Heart of Software* (Boston: Addison-Wesley Professional, 2003). For information about how you can use domain-driven design with the strangler fig pattern, see [Modernizing legacy Microsoft ASP.NET \(ASMX\) web services incrementally by using containers and Amazon API Gateway](#).

## DR

See [disaster recovery](#).

## drift detection

Tracking deviations from a baselined configuration. For example, you can use AWS CloudFormation to [detect drift in system resources](#), or you can use AWS Control Tower to [detect changes in your landing zone](#) that might affect compliance with governance requirements.

## DVSM

See [development value stream mapping](#).

## E

### EDA

See [exploratory data analysis](#).

### EDI

See [electronic data interchange](#).

### edge computing

The technology that increases the computing power for smart devices at the edges of an IoT network. When compared with [cloud computing](#), edge computing can reduce communication latency and improve response time.

### electronic data interchange (EDI)

The automated exchange of business documents between organizations. For more information, see [What is Electronic Data Interchange](#).

### encryption

A computing process that transforms plaintext data, which is human-readable, into ciphertext.

### encryption key

A cryptographic string of randomized bits that is generated by an encryption algorithm. Keys can vary in length, and each key is designed to be unpredictable and unique.

### endianness

The order in which bytes are stored in computer memory. Big-endian systems store the most significant byte first. Little-endian systems store the least significant byte first.

### endpoint

See [service endpoint](#).

### endpoint service

A service that you can host in a virtual private cloud (VPC) to share with other users. You can create an endpoint service with AWS PrivateLink and grant permissions to other AWS accounts or to AWS Identity and Access Management (IAM) principals. These accounts or principals can connect to your endpoint service privately by creating interface VPC endpoints. For more

information, see [Create an endpoint service](#) in the Amazon Virtual Private Cloud (Amazon VPC) documentation.

## enterprise resource planning (ERP)

A system that automates and manages key business processes (such as accounting, [MES](#), and project management) for an enterprise.

## envelope encryption

The process of encrypting an encryption key with another encryption key. For more information, see [Envelope encryption](#) in the AWS Key Management Service (AWS KMS) documentation.

## environment

An instance of a running application. The following are common types of environments in cloud computing:

- development environment – An instance of a running application that is available only to the core team responsible for maintaining the application. Development environments are used to test changes before promoting them to upper environments. This type of environment is sometimes referred to as a *test environment*.
- lower environments – All development environments for an application, such as those used for initial builds and tests.
- production environment – An instance of a running application that end users can access. In a CI/CD pipeline, the production environment is the last deployment environment.
- upper environments – All environments that can be accessed by users other than the core development team. This can include a production environment, preproduction environments, and environments for user acceptance testing.

## epic

In agile methodologies, functional categories that help organize and prioritize your work. Epics provide a high-level description of requirements and implementation tasks. For example, AWS CAF security epics include identity and access management, detective controls, infrastructure security, data protection, and incident response. For more information about epics in the AWS migration strategy, see the [program implementation guide](#).

## ERP

See [enterprise resource planning](#).

## exploratory data analysis (EDA)

The process of analyzing a dataset to understand its main characteristics. You collect or aggregate data and then perform initial investigations to find patterns, detect anomalies, and check assumptions. EDA is performed by calculating summary statistics and creating data visualizations.

## F

### fact table

The central table in a [star schema](#). It stores quantitative data about business operations. Typically, a fact table contains two types of columns: those that contain measures and those that contain a foreign key to a dimension table.

### fail fast

A philosophy that uses frequent and incremental testing to reduce the development lifecycle. It is a critical part of an agile approach.

### fault isolation boundary

In the AWS Cloud, a boundary such as an Availability Zone, AWS Region, control plane, or data plane that limits the effect of a failure and helps improve the resilience of workloads. For more information, see [AWS Fault Isolation Boundaries](#).

### feature branch

See [branch](#).

### features

The input data that you use to make a prediction. For example, in a manufacturing context, features could be images that are periodically captured from the manufacturing line.

### feature importance

How significant a feature is for a model's predictions. This is usually expressed as a numerical score that can be calculated through various techniques, such as Shapley Additive Explanations (SHAP) and integrated gradients. For more information, see [Machine learning model interpretability with AWS](#).



## feature transformation

To optimize data for the ML process, including enriching data with additional sources, scaling values, or extracting multiple sets of information from a single data field. This enables the ML model to benefit from the data. For example, if you break down the "2021-05-27 00:15:37" date into "2021", "May", "Thu", and "15", you can help the learning algorithm learn nuanced patterns associated with different data components.

## few-shot prompting

Providing an [LLM](#) with a small number of examples that demonstrate the task and desired output before asking it to perform a similar task. This technique is an application of in-context learning, where models learn from examples (*shots*) that are embedded in prompts. Few-shot prompting can be effective for tasks that require specific formatting, reasoning, or domain knowledge. See also [zero-shot prompting](#).

## FGAC

See [fine-grained access control](#).

## fine-grained access control (FGAC)

The use of multiple conditions to allow or deny an access request.

## flash-cut migration

A database migration method that uses continuous data replication through [change data capture](#) to migrate data in the shortest time possible, instead of using a phased approach. The objective is to keep downtime to a minimum.

## FM

See [foundation model](#).

## foundation model (FM)

A large deep-learning neural network that has been training on massive datasets of generalized and unlabeled data. FMs are capable of performing a wide variety of general tasks, such as understanding language, generating text and images, and conversing in natural language. For more information, see [What are Foundation Models](#).

# G

## generative AI

A subset of [AI](#) models that have been trained on large amounts of data and that can use a simple text prompt to create new content and artifacts, such as images, videos, text, and audio. For more information, see [What is Generative AI](#).

## geo blocking

See [geographic restrictions](#).

## geographic restrictions (geo blocking)

In Amazon CloudFront, an option to prevent users in specific countries from accessing content distributions. You can use an allow list or block list to specify approved and banned countries. For more information, see [Restricting the geographic distribution of your content](#) in the CloudFront documentation.

## Gitflow workflow

An approach in which lower and upper environments use different branches in a source code repository. The Gitflow workflow is considered legacy, and the [trunk-based workflow](#) is the modern, preferred approach.

## golden image

A snapshot of a system or software that is used as a template to deploy new instances of that system or software. For example, in manufacturing, a golden image can be used to provision software on multiple devices and helps improve speed, scalability, and productivity in device manufacturing operations.

## greenfield strategy

The absence of existing infrastructure in a new environment. When adopting a greenfield strategy for a system architecture, you can select all new technologies without the restriction of compatibility with existing infrastructure, also known as [brownfield](#). If you are expanding the existing infrastructure, you might blend brownfield and greenfield strategies.

## guardrail

A high-level rule that helps govern resources, policies, and compliance across organizational units (OUs). *Preventive guardrails* enforce policies to ensure alignment to compliance standards. They are implemented by using service control policies and IAM permissions boundaries.

*Detective guardrails* detect policy violations and compliance issues, and generate alerts for remediation. They are implemented by using AWS Config, AWS Security Hub, Amazon GuardDuty, AWS Trusted Advisor, Amazon Inspector, and custom AWS Lambda checks.

## H

### HA

See [high availability](#).

### heterogeneous database migration

Migrating your source database to a target database that uses a different database engine (for example, Oracle to Amazon Aurora). Heterogeneous migration is typically part of a re-architecting effort, and converting the schema can be a complex task. [AWS provides AWS SCT](#) that helps with schema conversions.

### high availability (HA)

The ability of a workload to operate continuously, without intervention, in the event of challenges or disasters. HA systems are designed to automatically fail over, consistently deliver high-quality performance, and handle different loads and failures with minimal performance impact.

### historian modernization

An approach used to modernize and upgrade operational technology (OT) systems to better serve the needs of the manufacturing industry. A *historian* is a type of database that is used to collect and store data from various sources in a factory.

### holdout data

A portion of historical, labeled data that is withheld from a dataset that is used to train a [machine learning](#) model. You can use holdout data to evaluate the model performance by comparing the model predictions against the holdout data.

### homogeneous database migration

Migrating your source database to a target database that shares the same database engine (for example, Microsoft SQL Server to Amazon RDS for SQL Server). Homogeneous migration is typically part of a rehosting or replatforming effort. You can use native database utilities to migrate the schema.

## hot data

Data that is frequently accessed, such as real-time data or recent translational data. This data typically requires a high-performance storage tier or class to provide fast query responses.

## hotfix

An urgent fix for a critical issue in a production environment. Due to its urgency, a hotfix is usually made outside of the typical DevOps release workflow.

## hypercare period

Immediately following cutover, the period of time when a migration team manages and monitors the migrated applications in the cloud in order to address any issues. Typically, this period is 1–4 days in length. At the end of the hypercare period, the migration team typically transfers responsibility for the applications to the cloud operations team.

## I

### IaC

See [infrastructure as code](#).

### identity-based policy

A policy attached to one or more IAM principals that defines their permissions within the AWS Cloud environment.

### idle application

An application that has an average CPU and memory usage between 5 and 20 percent over a period of 90 days. In a migration project, it is common to retire these applications or retain them on premises.

## IIoT

See [Industrial Internet of Things](#).

### immutable infrastructure

A model that deploys new infrastructure for production workloads instead of updating, patching, or modifying the existing infrastructure. Immutable infrastructures are inherently more consistent, reliable, and predictable than [mutable infrastructure](#). For more information, see the [Deploy using immutable infrastructure](#) best practice in the AWS Well-Architected Framework.

## inbound (ingress) VPC

In an AWS multi-account architecture, a VPC that accepts, inspects, and routes network connections from outside an application. The [AWS Security Reference Architecture](#) recommends setting up your Network account with inbound, outbound, and inspection VPCs to protect the two-way interface between your application and the broader internet.

## incremental migration

A cutover strategy in which you migrate your application in small parts instead of performing a single, full cutover. For example, you might move only a few microservices or users to the new system initially. After you verify that everything is working properly, you can incrementally move additional microservices or users until you can decommission your legacy system. This strategy reduces the risks associated with large migrations.

## Industry 4.0

A term that was introduced by [Klaus Schwab](#) in 2016 to refer to the modernization of manufacturing processes through advances in connectivity, real-time data, automation, analytics, and AI/ML.

## infrastructure

All of the resources and assets contained within an application's environment.

## infrastructure as code (IaC)

The process of provisioning and managing an application's infrastructure through a set of configuration files. IaC is designed to help you centralize infrastructure management, standardize resources, and scale quickly so that new environments are repeatable, reliable, and consistent.

## industrial Internet of Things (IIoT)

The use of internet-connected sensors and devices in the industrial sectors, such as manufacturing, energy, automotive, healthcare, life sciences, and agriculture. For more information, see [Building an industrial Internet of Things \(IIoT\) digital transformation strategy](#).

## inspection VPC

In an AWS multi-account architecture, a centralized VPC that manages inspections of network traffic between VPCs (in the same or different AWS Regions), the internet, and on-premises networks. The [AWS Security Reference Architecture](#) recommends setting up your Network account with inbound, outbound, and inspection VPCs to protect the two-way interface between your application and the broader internet.

## Internet of Things (IoT)

The network of connected physical objects with embedded sensors or processors that communicate with other devices and systems through the internet or over a local communication network. For more information, see [What is IoT?](#)

## interpretability

A characteristic of a machine learning model that describes the degree to which a human can understand how the model's predictions depend on its inputs. For more information, see [Machine learning model interpretability with AWS](#).

## IoT

See [Internet of Things](#).

## IT information library (ITIL)

A set of best practices for delivering IT services and aligning these services with business requirements. ITIL provides the foundation for ITSM.

## IT service management (ITSM)

Activities associated with designing, implementing, managing, and supporting IT services for an organization. For information about integrating cloud operations with ITSM tools, see the [operations integration guide](#).

## ITIL

See [IT information library](#).

## ITSM

See [IT service management](#).

# L

## label-based access control (LBAC)

An implementation of mandatory access control (MAC) where the users and the data itself are each explicitly assigned a security label value. The intersection between the user security label and data security label determines which rows and columns can be seen by the user.

## landing zone

A landing zone is a well-architected, multi-account AWS environment that is scalable and secure. This is a starting point from which your organizations can quickly launch and deploy workloads and applications with confidence in their security and infrastructure environment. For more information about landing zones, see [Setting up a secure and scalable multi-account AWS environment](#).

## large language model (LLM)

A deep learning [AI](#) model that is pretrained on a vast amount of data. An LLM can perform multiple tasks, such as answering questions, summarizing documents, translating text into other languages, and completing sentences. For more information, see [What are LLMs](#).

## large migration

A migration of 300 or more servers.

## LBAC

See [label-based access control](#).

## least privilege

The security best practice of granting the minimum permissions required to perform a task. For more information, see [Apply least-privilege permissions](#) in the IAM documentation.

## lift and shift

See [7 Rs](#).

## little-endian system

A system that stores the least significant byte first. See also [endianness](#).

## LLM

See [large language model](#).

## lower environments

See [environment](#).

# M

## machine learning (ML)

A type of artificial intelligence that uses algorithms and techniques for pattern recognition and learning. ML analyzes and learns from recorded data, such as Internet of Things (IoT) data, to generate a statistical model based on patterns. For more information, see [Machine Learning](#).

## main branch

See [branch](#).

## malware

Software that is designed to compromise computer security or privacy. Malware might disrupt computer systems, leak sensitive information, or gain unauthorized access. Examples of malware include viruses, worms, ransomware, Trojan horses, spyware, and keyloggers.

## managed services

AWS services for which AWS operates the infrastructure layer, the operating system, and platforms, and you access the endpoints to store and retrieve data. Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB are examples of managed services. These are also known as *abstracted services*.

## manufacturing execution system (MES)

A software system for tracking, monitoring, documenting, and controlling production processes that convert raw materials to finished products on the shop floor.

## MAP

See [Migration Acceleration Program](#).

## mechanism

A complete process in which you create a tool, drive adoption of the tool, and then inspect the results in order to make adjustments. A mechanism is a cycle that reinforces and improves itself as it operates. For more information, see [Building mechanisms](#) in the AWS Well-Architected Framework.

## member account

All AWS accounts other than the management account that are part of an organization in AWS Organizations. An account can be a member of only one organization at a time.



## MES

See [manufacturing execution system](#).

## Message Queuing Telemetry Transport (MQTT)

A lightweight, machine-to-machine (M2M) communication protocol, based on the [publish/subscribe](#) pattern, for resource-constrained [IoT](#) devices.

## microservice

A small, independent service that communicates over well-defined APIs and is typically owned by small, self-contained teams. For example, an insurance system might include microservices that map to business capabilities, such as sales or marketing, or subdomains, such as purchasing, claims, or analytics. The benefits of microservices include agility, flexible scaling, easy deployment, reusable code, and resilience. For more information, see [Integrating microservices by using AWS serverless services](#).

## microservices architecture

An approach to building an application with independent components that run each application process as a microservice. These microservices communicate through a well-defined interface by using lightweight APIs. Each microservice in this architecture can be updated, deployed, and scaled to meet demand for specific functions of an application. For more information, see [Implementing microservices on AWS](#).

## Migration Acceleration Program (MAP)

An AWS program that provides consulting support, training, and services to help organizations build a strong operational foundation for moving to the cloud, and to help offset the initial cost of migrations. MAP includes a migration methodology for executing legacy migrations in a methodical way and a set of tools to automate and accelerate common migration scenarios.

## migration at scale

The process of moving the majority of the application portfolio to the cloud in waves, with more applications moved at a faster rate in each wave. This phase uses the best practices and lessons learned from the earlier phases to implement a *migration factory* of teams, tools, and processes to streamline the migration of workloads through automation and agile delivery. This is the third phase of the [AWS migration strategy](#).

## migration factory

Cross-functional teams that streamline the migration of workloads through automated, agile approaches. Migration factory teams typically include operations, business analysts and owners,

migration engineers, developers, and DevOps professionals working in sprints. Between 20 and 50 percent of an enterprise application portfolio consists of repeated patterns that can be optimized by a factory approach. For more information, see the [discussion of migration factories](#) and the [Cloud Migration Factory guide](#) in this content set.

## migration metadata

The information about the application and server that is needed to complete the migration. Each migration pattern requires a different set of migration metadata. Examples of migration metadata include the target subnet, security group, and AWS account.

## migration pattern

A repeatable migration task that details the migration strategy, the migration destination, and the migration application or service used. Example: Rehost migration to Amazon EC2 with AWS Application Migration Service.

## Migration Portfolio Assessment (MPA)

An online tool that provides information for validating the business case for migrating to the AWS Cloud. MPA provides detailed portfolio assessment (server right-sizing, pricing, TCO comparisons, migration cost analysis) as well as migration planning (application data analysis and data collection, application grouping, migration prioritization, and wave planning). The [MPA tool](#) (requires login) is available free of charge to all AWS consultants and APN Partner consultants.

## Migration Readiness Assessment (MRA)

The process of gaining insights about an organization's cloud readiness status, identifying strengths and weaknesses, and building an action plan to close identified gaps, using the AWS CAF. For more information, see the [migration readiness guide](#). MRA is the first phase of the [AWS migration strategy](#).

## migration strategy

The approach used to migrate a workload to the AWS Cloud. For more information, see the [7 Rs](#) entry in this glossary and see [Mobilize your organization to accelerate large-scale migrations](#).

## ML

See [machine learning](#).

## modernization

Transforming an outdated (legacy or monolithic) application and its infrastructure into an agile, elastic, and highly available system in the cloud to reduce costs, gain efficiencies, and take advantage of innovations. For more information, see [Strategy for modernizing applications in the AWS Cloud](#).

## modernization readiness assessment

An evaluation that helps determine the modernization readiness of an organization's applications; identifies benefits, risks, and dependencies; and determines how well the organization can support the future state of those applications. The outcome of the assessment is a blueprint of the target architecture, a roadmap that details development phases and milestones for the modernization process, and an action plan for addressing identified gaps. For more information, see [Evaluating modernization readiness for applications in the AWS Cloud](#).

## monolithic applications (monoliths)

Applications that run as a single service with tightly coupled processes. Monolithic applications have several drawbacks. If one application feature experiences a spike in demand, the entire architecture must be scaled. Adding or improving a monolithic application's features also becomes more complex when the code base grows. To address these issues, you can use a microservices architecture. For more information, see [Decomposing monoliths into microservices](#).

## MPA

See [Migration Portfolio Assessment](#).

## MQTT

See [Message Queuing Telemetry Transport](#).

## multiclass classification

A process that helps generate predictions for multiple classes (predicting one of more than two outcomes). For example, an ML model might ask "Is this product a book, car, or phone?" or "Which product category is most interesting to this customer?"

## mutable infrastructure

A model that updates and modifies the existing infrastructure for production workloads. For improved consistency, reliability, and predictability, the AWS Well-Architected Framework recommends the use of [immutable infrastructure](#) as a best practice.

# O

## OAC

See [origin access control](#).

## OAI

See [origin access identity](#).

## OCM

See [organizational change management](#).

## offline migration

A migration method in which the source workload is taken down during the migration process. This method involves extended downtime and is typically used for small, non-critical workloads.

## OI

See [operations integration](#).

## OLA

See [operational-level agreement](#).

## online migration

A migration method in which the source workload is copied to the target system without being taken offline. Applications that are connected to the workload can continue to function during the migration. This method involves zero to minimal downtime and is typically used for critical production workloads.

## OPC-UA

See [Open Process Communications - Unified Architecture](#).

## Open Process Communications - Unified Architecture (OPC-UA)

A machine-to-machine (M2M) communication protocol for industrial automation. OPC-UA provides an interoperability standard with data encryption, authentication, and authorization schemes.

## operational-level agreement (OLA)

An agreement that clarifies what functional IT groups promise to deliver to each other, to support a service-level agreement (SLA).

## operational readiness review (ORR)

A checklist of questions and associated best practices that help you understand, evaluate, prevent, or reduce the scope of incidents and possible failures. For more information, see [Operational Readiness Reviews \(ORR\)](#) in the AWS Well-Architected Framework.

## operational technology (OT)

Hardware and software systems that work with the physical environment to control industrial operations, equipment, and infrastructure. In manufacturing, the integration of OT and information technology (IT) systems is a key focus for [Industry 4.0](#) transformations.

## operations integration (OI)

The process of modernizing operations in the cloud, which involves readiness planning, automation, and integration. For more information, see the [operations integration guide](#).

## organization trail

A trail that's created by AWS CloudTrail that logs all events for all AWS accounts in an organization in AWS Organizations. This trail is created in each AWS account that's part of the organization and tracks the activity in each account. For more information, see [Creating a trail for an organization](#) in the CloudTrail documentation.

## organizational change management (OCM)

A framework for managing major, disruptive business transformations from a people, culture, and leadership perspective. OCM helps organizations prepare for, and transition to, new systems and strategies by accelerating change adoption, addressing transitional issues, and driving cultural and organizational changes. In the AWS migration strategy, this framework is called *people acceleration*, because of the speed of change required in cloud adoption projects. For more information, see the [OCM guide](#).

## origin access control (OAC)

In CloudFront, an enhanced option for restricting access to secure your Amazon Simple Storage Service (Amazon S3) content. OAC supports all S3 buckets in all AWS Regions, server-side encryption with AWS KMS (SSE-KMS), and dynamic PUT and DELETE requests to the S3 bucket.

## origin access identity (OAI)

In CloudFront, an option for restricting access to secure your Amazon S3 content. When you use OAI, CloudFront creates a principal that Amazon S3 can authenticate with. Authenticated principals can access content in an S3 bucket only through a specific CloudFront distribution. See also [OAC](#), which provides more granular and enhanced access control.

## ORR

See [operational readiness review](#).

## OT

See [operational technology](#).

## outbound (egress) VPC

In an AWS multi-account architecture, a VPC that handles network connections that are initiated from within an application. The [AWS Security Reference Architecture](#) recommends setting up your Network account with inbound, outbound, and inspection VPCs to protect the two-way interface between your application and the broader internet.

## P

### permissions boundary

An IAM management policy that is attached to IAM principals to set the maximum permissions that the user or role can have. For more information, see [Permissions boundaries](#) in the IAM documentation.

### personally identifiable information (PII)

Information that, when viewed directly or paired with other related data, can be used to reasonably infer the identity of an individual. Examples of PII include names, addresses, and contact information.

## PII

See [personally identifiable information](#).

## playbook

A set of predefined steps that capture the work associated with migrations, such as delivering core operations functions in the cloud. A playbook can take the form of scripts, automated runbooks, or a summary of processes or steps required to operate your modernized environment.

## PLC

See [programmable logic controller](#).

## PLM

See [product lifecycle management](#).

### policy

An object that can define permissions (see [identity-based policy](#)), specify access conditions (see [resource-based policy](#)), or define the maximum permissions for all accounts in an organization in AWS Organizations (see [service control policy](#)).

### polyglot persistence

Independently choosing a microservice's data storage technology based on data access patterns and other requirements. If your microservices have the same data storage technology, they can encounter implementation challenges or experience poor performance. Microservices are more easily implemented and achieve better performance and scalability if they use the data store best adapted to their requirements. For more information, see [Enabling data persistence in microservices](#).

### portfolio assessment

A process of discovering, analyzing, and prioritizing the application portfolio in order to plan the migration. For more information, see [Evaluating migration readiness](#).

### predicate

A query condition that returns true or false, commonly located in a WHERE clause.

### predicate pushdown

A database query optimization technique that filters the data in the query before transfer. This reduces the amount of data that must be retrieved and processed from the relational database, and it improves query performance.

### preventative control

A security control that is designed to prevent an event from occurring. These controls are a first line of defense to help prevent unauthorized access or unwanted changes to your network. For more information, see [Preventative controls](#) in *Implementing security controls on AWS*.

### principal

An entity in AWS that can perform actions and access resources. This entity is typically a root user for an AWS account, an IAM role, or a user. For more information, see *Principal* in [Roles terms and concepts](#) in the IAM documentation.

## privacy by design

A system engineering approach that takes privacy into account through the whole development process.

## private hosted zones

A container that holds information about how you want Amazon Route 53 to respond to DNS queries for a domain and its subdomains within one or more VPCs. For more information, see [Working with private hosted zones](#) in the Route 53 documentation.

## proactive control

A [security control](#) designed to prevent the deployment of noncompliant resources. These controls scan resources before they are provisioned. If the resource is not compliant with the control, then it isn't provisioned. For more information, see the [Controls reference guide](#) in the AWS Control Tower documentation and see [Proactive controls](#) in *Implementing security controls on AWS*.

## product lifecycle management (PLM)

The management of data and processes for a product throughout its entire lifecycle, from design, development, and launch, through growth and maturity, to decline and removal.

## production environment

See [environment](#).

## programmable logic controller (PLC)

In manufacturing, a highly reliable, adaptable computer that monitors machines and automates manufacturing processes.

## prompt chaining

Using the output of one [LLM](#) prompt as the input for the next prompt to generate better responses. This technique is used to break down a complex task into subtasks, or to iteratively refine or expand a preliminary response. It helps improve the accuracy and relevance of a model's responses and allows for more granular, personalized results.

## pseudonymization

The process of replacing personal identifiers in a dataset with placeholder values. Pseudonymization can help protect personal privacy. Pseudonymized data is still considered to be personal data.



## publish/subscribe (pub/sub)

A pattern that enables asynchronous communications among microservices to improve scalability and responsiveness. For example, in a microservices-based [MES](#), a microservice can publish event messages to a channel that other microservices can subscribe to. The system can add new microservices without changing the publishing service.

## Q

### query plan

A series of steps, like instructions, that are used to access the data in a SQL relational database system.

### query plan regression

When a database service optimizer chooses a less optimal plan than it did before a given change to the database environment. This can be caused by changes to statistics, constraints, environment settings, query parameter bindings, and updates to the database engine.

## R

### RACI matrix

See [responsible, accountable, consulted, informed \(RACI\)](#).

### RAG

See [Retrieval Augmented Generation](#).

### ransomware

A malicious software that is designed to block access to a computer system or data until a payment is made.

### RASCI matrix

See [responsible, accountable, consulted, informed \(RACI\)](#).

### RCAC

See [row and column access control](#).

## read replica

A copy of a database that's used for read-only purposes. You can route queries to the read replica to reduce the load on your primary database.

## re-architect

See [7 Rs](#).

## recovery point objective (RPO)

The maximum acceptable amount of time since the last data recovery point. This determines what is considered an acceptable loss of data between the last recovery point and the interruption of service.

## recovery time objective (RTO)

The maximum acceptable delay between the interruption of service and restoration of service.

## refactor

See [7 Rs](#).

## Region

A collection of AWS resources in a geographic area. Each AWS Region is isolated and independent of the others to provide fault tolerance, stability, and resilience. For more information, see [Specify which AWS Regions your account can use](#).

## regression

An ML technique that predicts a numeric value. For example, to solve the problem of "What price will this house sell for?" an ML model could use a linear regression model to predict a house's sale price based on known facts about the house (for example, the square footage).

## rehost

See [7 Rs](#).

## release

In a deployment process, the act of promoting changes to a production environment.

## relocate

See [7 Rs](#).

## replatform

See [7 Rs](#).

## repurchase

See [7 Rs](#).

## resiliency

An application's ability to resist or recover from disruptions. [High availability](#) and [disaster recovery](#) are common considerations when planning for resiliency in the AWS Cloud. For more information, see [AWS Cloud Resilience](#).

## resource-based policy

A policy attached to a resource, such as an Amazon S3 bucket, an endpoint, or an encryption key. This type of policy specifies which principals are allowed access, supported actions, and any other conditions that must be met.

## responsible, accountable, consulted, informed (RACI) matrix

A matrix that defines the roles and responsibilities for all parties involved in migration activities and cloud operations. The matrix name is derived from the responsibility types defined in the matrix: responsible (R), accountable (A), consulted (C), and informed (I). The support (S) type is optional. If you include support, the matrix is called a *RASCI matrix*, and if you exclude it, it's called a *RACI matrix*.

## responsive control

A security control that is designed to drive remediation of adverse events or deviations from your security baseline. For more information, see [Responsive controls](#) in *Implementing security controls on AWS*.

## retain

See [7 Rs](#).

## retire

See [7 Rs](#).

## Retrieval Augmented Generation (RAG)

A [generative AI](#) technology in which an [LLM](#) references an authoritative data source that is outside of its training data sources before generating a response. For example, a RAG model might perform a semantic search of an organization's knowledge base or custom data. For more information, see [What is RAG](#).

## rotation

The process of periodically updating a [secret](#) to make it more difficult for an attacker to access the credentials.

## row and column access control (RCAC)

The use of basic, flexible SQL expressions that have defined access rules. RCAC consists of row permissions and column masks.

## RPO

See [recovery point objective](#).

## RTO

See [recovery time objective](#).

## runbook

A set of manual or automated procedures required to perform a specific task. These are typically built to streamline repetitive operations or procedures with high error rates.

# S

## SAML 2.0

An open standard that many identity providers (IdPs) use. This feature enables federated single sign-on (SSO), so users can log into the AWS Management Console or call the AWS API operations without you having to create user in IAM for everyone in your organization. For more information about SAML 2.0-based federation, see [About SAML 2.0-based federation](#) in the IAM documentation.

## SCADA

See [supervisory control and data acquisition](#).

## SCP

See [service control policy](#).

## secret

In AWS Secrets Manager, confidential or restricted information, such as a password or user credentials, that you store in encrypted form. It consists of the secret value and its metadata.

The secret value can be binary, a single string, or multiple strings. For more information, see [What's in a Secrets Manager secret?](#) in the Secrets Manager documentation.

## security by design

A system engineering approach that takes security into account through the whole development process.

## security control

A technical or administrative guardrail that prevents, detects, or reduces the ability of a threat actor to exploit a security vulnerability. There are four primary types of security controls: [preventative](#), [detective](#), [responsive](#), and [proactive](#).

## security hardening

The process of reducing the attack surface to make it more resistant to attacks. This can include actions such as removing resources that are no longer needed, implementing the security best practice of granting least privilege, or deactivating unnecessary features in configuration files.

## security information and event management (SIEM) system

Tools and services that combine security information management (SIM) and security event management (SEM) systems. A SIEM system collects, monitors, and analyzes data from servers, networks, devices, and other sources to detect threats and security breaches, and to generate alerts.

## security response automation

A predefined and programmed action that is designed to automatically respond to or remediate a security event. These automations serve as [detective](#) or [responsive](#) security controls that help you implement AWS security best practices. Examples of automated response actions include modifying a VPC security group, patching an Amazon EC2 instance, or rotating credentials.

## server-side encryption

Encryption of data at its destination, by the AWS service that receives it.

## service control policy (SCP)

A policy that provides centralized control over permissions for all accounts in an organization in AWS Organizations. SCPs define guardrails or set limits on actions that an administrator can delegate to users or roles. You can use SCPs as allow lists or deny lists, to specify which services or actions are permitted or prohibited. For more information, see [Service control policies](#) in the AWS Organizations documentation.

## service endpoint

The URL of the entry point for an AWS service. You can use the endpoint to connect programmatically to the target service. For more information, see [AWS service endpoints](#) in *AWS General Reference*.

## service-level agreement (SLA)

An agreement that clarifies what an IT team promises to deliver to their customers, such as service uptime and performance.

## service-level indicator (SLI)

A measurement of a performance aspect of a service, such as its error rate, availability, or throughput.

## service-level objective (SLO)

A target metric that represents the health of a service, as measured by a [service-level indicator](#).

## shared responsibility model

A model describing the responsibility you share with AWS for cloud security and compliance. AWS is responsible for security *of* the cloud, whereas you are responsible for security *in* the cloud. For more information, see [Shared responsibility model](#).

## SIEM

See [security information and event management system](#).

## single point of failure (SPOF)

A failure in a single, critical component of an application that can disrupt the system.

## SLA

See [service-level agreement](#).

## SLI

See [service-level indicator](#).

## SLO

See [service-level objective](#).

## split-and-seed model

A pattern for scaling and accelerating modernization projects. As new features and product releases are defined, the core team splits up to create new product teams. This helps scale your

organization's capabilities and services, improves developer productivity, and supports rapid innovation. For more information, see [Phased approach to modernizing applications in the AWS Cloud](#).

## SPOF

See [single point of failure](#).

## star schema

A database organizational structure that uses one large fact table to store transactional or measured data and uses one or more smaller dimensional tables to store data attributes. This structure is designed for use in a [data warehouse](#) or for business intelligence purposes.

## strangler fig pattern

An approach to modernizing monolithic systems by incrementally rewriting and replacing system functionality until the legacy system can be decommissioned. This pattern uses the analogy of a fig vine that grows into an established tree and eventually overcomes and replaces its host. The pattern was [introduced by Martin Fowler](#) as a way to manage risk when rewriting monolithic systems. For an example of how to apply this pattern, see [Modernizing legacy Microsoft ASP.NET \(ASMX\) web services incrementally by using containers and Amazon API Gateway](#).

## subnet

A range of IP addresses in your VPC. A subnet must reside in a single Availability Zone.

## supervisory control and data acquisition (SCADA)

In manufacturing, a system that uses hardware and software to monitor physical assets and production operations.

## symmetric encryption

An encryption algorithm that uses the same key to encrypt and decrypt the data.

## synthetic testing

Testing a system in a way that simulates user interactions to detect potential issues or to monitor performance. You can use [Amazon CloudWatch Synthetics](#) to create these tests.

## system prompt

A technique for providing context, instructions, or guidelines to an [LLM](#) to direct its behavior. System prompts help set context and establish rules for interactions with users.

# T

## tags

Key-value pairs that act as metadata for organizing your AWS resources. Tags can help you manage, identify, organize, search for, and filter resources. For more information, see [Tagging your AWS resources](#).

## target variable

The value that you are trying to predict in supervised ML. This is also referred to as an *outcome variable*. For example, in a manufacturing setting the target variable could be a product defect.

## task list

A tool that is used to track progress through a runbook. A task list contains an overview of the runbook and a list of general tasks to be completed. For each general task, it includes the estimated amount of time required, the owner, and the progress.

## test environment

See [environment](#).

## training

To provide data for your ML model to learn from. The training data must contain the correct answer. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that you want to predict). It outputs an ML model that captures these patterns. You can then use the ML model to make predictions on new data for which you don't know the target.

## transit gateway

A network transit hub that you can use to interconnect your VPCs and on-premises networks. For more information, see [What is a transit gateway](#) in the AWS Transit Gateway documentation.

## trunk-based workflow

An approach in which developers build and test features locally in a feature branch and then merge those changes into the main branch. The main branch is then built to the development, preproduction, and production environments, sequentially.



## trusted access

Granting permissions to a service that you specify to perform tasks in your organization in AWS Organizations and in its accounts on your behalf. The trusted service creates a service-linked role in each account, when that role is needed, to perform management tasks for you. For more information, see [Using AWS Organizations with other AWS services](#) in the AWS Organizations documentation.

## tuning

To change aspects of your training process to improve the ML model's accuracy. For example, you can train the ML model by generating a labeling set, adding labels, and then repeating these steps several times under different settings to optimize the model.

## two-pizza team

A small DevOps team that you can feed with two pizzas. A two-pizza team size ensures the best possible opportunity for collaboration in software development.

# U

## uncertainty

A concept that refers to imprecise, incomplete, or unknown information that can undermine the reliability of predictive ML models. There are two types of uncertainty: *Epistemic uncertainty* is caused by limited, incomplete data, whereas *aleatoric uncertainty* is caused by the noise and randomness inherent in the data. For more information, see the [Quantifying uncertainty in deep learning systems](#) guide.

## undifferentiated tasks

Also known as *heavy lifting*, work that is necessary to create and operate an application but that doesn't provide direct value to the end user or provide competitive advantage. Examples of undifferentiated tasks include procurement, maintenance, and capacity planning.

## upper environments

See [environment](#).

## V

### vacuuming

A database maintenance operation that involves cleaning up after incremental updates to reclaim storage and improve performance.

### version control

Processes and tools that track changes, such as changes to source code in a repository.

### VPC peering

A connection between two VPCs that allows you to route traffic by using private IP addresses. For more information, see [What is VPC peering](#) in the Amazon VPC documentation.

### vulnerability

A software or hardware flaw that compromises the security of the system.

## W

### warm cache

A buffer cache that contains current, relevant data that is frequently accessed. The database instance can read from the buffer cache, which is faster than reading from the main memory or disk.

### warm data

Data that is infrequently accessed. When querying this kind of data, moderately slow queries are typically acceptable.

### window function

A SQL function that performs a calculation on a group of rows that relate in some way to the current record. Window functions are useful for processing tasks, such as calculating a moving average or accessing the value of rows based on the relative position of the current row.

### workload

A collection of resources and code that delivers business value, such as a customer-facing application or backend process.

## workstream

Functional groups in a migration project that are responsible for a specific set of tasks. Each workstream is independent but supports the other workstreams in the project. For example, the portfolio workstream is responsible for prioritizing applications, wave planning, and collecting migration metadata. The portfolio workstream delivers these assets to the migration workstream, which then migrates the servers and applications.

## WORM

See [write once, read many](#).

## WQF

See [AWS Workload Qualification Framework](#).

## write once, read many (WORM)

A storage model that writes data a single time and prevents the data from being deleted or modified. Authorized users can read the data as many times as needed, but they cannot change it. This data storage infrastructure is considered [immutable](#).

# Z

## zero-day exploit

An attack, typically malware, that takes advantage of a [zero-day vulnerability](#).

## zero-day vulnerability

An unmitigated flaw or vulnerability in a production system. Threat actors can use this type of vulnerability to attack the system. Developers frequently become aware of the vulnerability as a result of the attack.

## zero-shot prompting

Providing an [LLM](#) with instructions for performing a task but no examples (*shots*) that can help guide it. The LLM must use its pre-trained knowledge to handle the task. The effectiveness of zero-shot prompting depends on the complexity of the task and the quality of the prompt. See also [few-shot prompting](#).

## zombie application

An application that has an average CPU and memory usage below 5 percent. In a migration project, it is common to retire these applications.