Preparing for the California Consumer Privacy Act

First Published July 2019

Updated November 3, 2021

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

Contents

Abstract

This document provides information to assist customers subject to the California Consumer Privacy Act (CCPA) as they accelerate their use of Amazon Web Services (AWS) cloud services.

Introduction

The California Consumer Privacy Act of 2018 (CCPA) grants "consumer[s] various rights with regard to personal information relating to the consumer that is held by a business" that is subject to the CCPA. Specifically, the CCPA grants "consumers" the right to request that a "business" disclose the categories and specific pieces of personal information collected about the consumer, the categories of sources from which that information is collected, the "business purposes" for collecting or selling the information, and the categories of third parties with which the information is shared.

The California Privacy Rights Act (CPRA), passed on November 3, 2020 and which will come into full effect on January 1, 2023, creates a new enforcement authority and will replace and build upon the CCPA. It adds, among other provisions, a right to rectification, right to restriction, and creates a category for "sensitive personal information." The CPRA states that businesses collecting personal information of consumers are required to clearly inform them when they employ automated decision-making technology. The CPRA strengthens punishments for breaches involving children's data. Any administrative fines are three times as much for children's personal information. The law also affects how consent is managed and obtained by regulated businesses, and includes provisions that afford parents greater control over the personal information of their children.

Additional enhancements include the obligation of companies to protect privacy rights of their employees and independent contractors. Businesses are explicitly obligated to protect their employees' data privacy, but there are some minor distinctions between their privacy and how consumers' privacy is handled. CPRA enables greater flexibility to the created enforcement agency to keep the privacy laws up to date over time, in an attempt to keep the law current and applicable. It gives the agency authority to prevent future attempts by businesses to circumvent or otherwise not comply with the CPRA.

This document begins with an overview of AWS security and compliance, then addresses the four main subsections of the CCPA: Data Collection, Zero Trust, Data Retrieval and Deletion, and Data Awareness.

Maintaining customer trust is an ongoing commitment at AWS. AWS strives to inform you of its privacy and data security policies, practices, and associated technologies. These commitments include:

- **Access** — As a customer, you maintain full control of your content and are responsible for configuring access to AWS services and resources. AWS provides an advanced set of access, encryption, and logging services, such as AWS Identity and Access Management (IAM), AWS Organizations, and AWS CloudTrail, to help you do this efficiently. AWS provides APIs that enable you to control access to any of the services you develop or deploy in AWS. AWS does not access or use your content for any purpose without your agreement. AWS never uses your content, or derives information or insights from it, for marketing or advertising purposes.
- **Storage** — You choose one or more AWS Regions in which to store your content, and select the type of storage that is used. You can replicate and back up your content to multiple AWS Regions, or to on-premises. AWS will not move or replicate your content outside of your chosen AWS Regions or on-premises area without your consent, except in cases where it is legally required, or is necessary to maintain the AWS services.
- **Security** — You choose how your content is secured. AWS offers you strong encryption for your content in transit and at rest, and provides you the ability to manage your own encryption keys. These features include:
    - **Data encryption capabilities** in AWS storage and database services, such as Amazon Elastic Block Store (EBS), Amazon Simple Storage Service (Amazon S3), Amazon Relational Database Service (Amazon RDS), Amazon DynamoDB, and Amazon Redshift. In addition, AWS API endpoints use Signature Version 4 (Sigv4) to add authentication to API requests sent by HTTP. This method uniquely authenticates and authorizes each and every signed API request, and provides fine-grained access controls.
    - **Flexible key management options**, including AWS Key Management Service (AWS KMS), which enable you to choose whether AWS manages your encryption keys for you, or whether you manage your keys yourself—giving you complete control over your encryption keys.

      You can employ server-side encryption (SSE) with Amazon Service Managed Keys, AWS Customer Managed Keys, or customer-provided encryption keys. This includes symmetric and asymmetric keys. You can also use AWS Certificate Manager to provision, manage and deploy public and private Secure Socket Later/Transport Later Security (SSL/TLS) certificates for use with AWS services and your internal connected resources. SSL/TLS certificates are used to secure network communications

and establish the identity of websites over the internet as well as resources on private networks.

- **Disclosure of customer content** — AWS does not disclose customer content unless required to do so to comply with a legally valid and binding order. If a government body sends AWS a demand for customer content, we will attempt to redirect the governmental body to request that data directly from the customer. If compelled to disclose customer content to a government body, we will give customers reasonable notice of the demand to allow the customer to seek protective order or other appropriate remedy unless AWS is legally prohibited from doing so.
- **Security Assurance** —AWS has a security assurance program that uses best practices for global privacy and data protection to help you operate securely within AWS, and to make the best use of our security control environment. These security protections and control processes are independently validated by multiple third-party independent assessments.

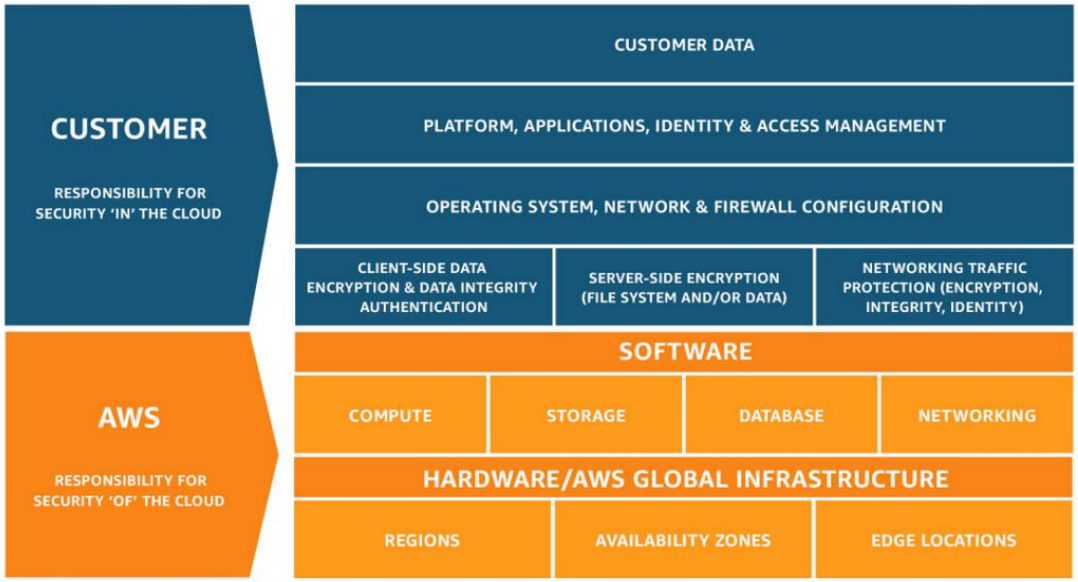
For guidance and best practices on building security policies and processes for your organization, see the *AWS Security Best Practices* whitepaper. For information on how AWS collects and uses personal information, see the AWS Privacy Notice.

Security and shared responsibility

Cloud security at AWS is the highest priority. Security and compliance are a shared responsibility between AWS and you, the customer. This shared model lessens your operational overhead because AWS operates, maintains, and controls the infrastructure, from the host operating system and virtualization layer down to the physical security of the facilities where the services run.

For infrastructure services, you assume responsibility for the management of the guest operating system (including installing updates and security patches), other associated application software, and the configuration of the AWS-provided security group firewall. Carefully consider the services that you choose to use, as your responsibilities vary depending on the services used, how those services are integrated into your IT environment, and applicable laws and regulations.

As shown in the following diagram, this differentiation of responsibility is commonly referred to as Security “of” the Cloud versus Security “in” the Cloud.

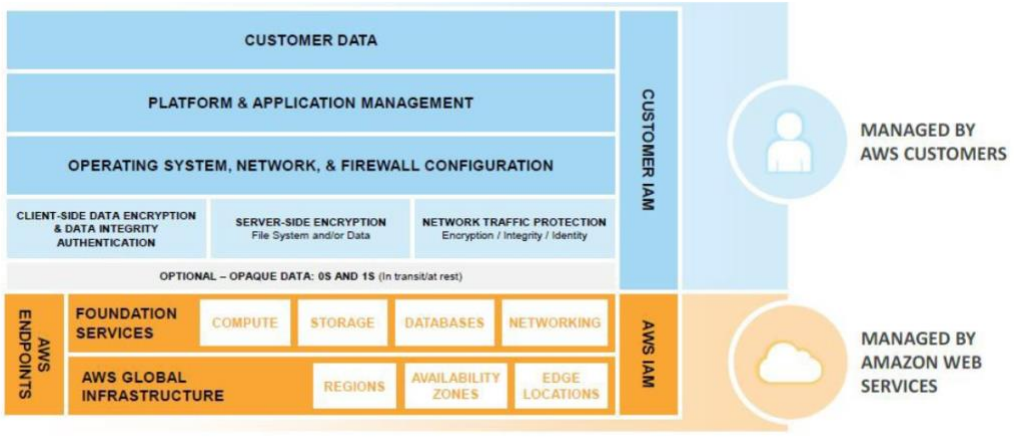


AWS shared responsibility model

As the following diagrams illustrate, as you progress up the application stack, the shared responsibility line moves up, with AWS taking over more of the responsibility. At higher levels in the application stack, the customer inherits all controls of the lower levels. Even at the highest level of the application stack, AWS does not exert any control over your data.

Shared responsibility model for infrastructure services

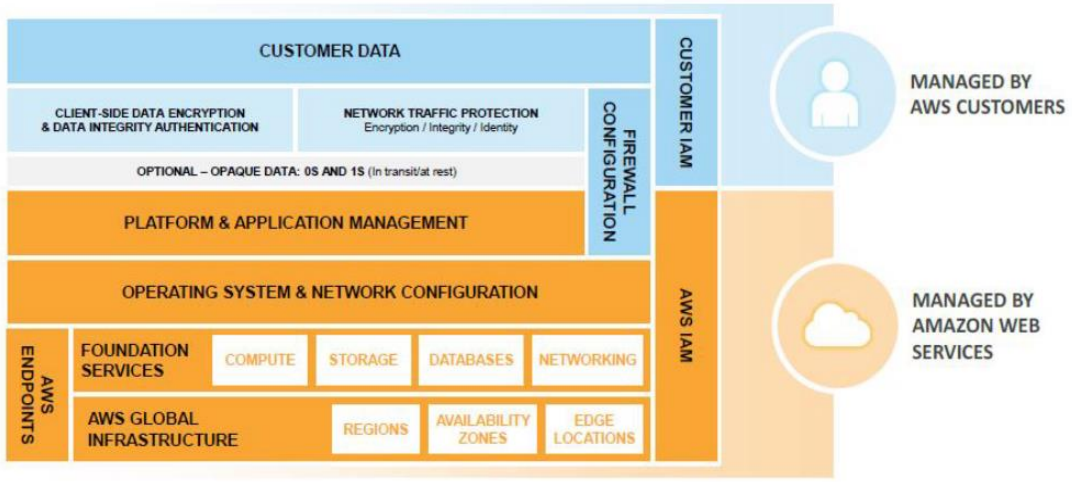
Infrastructure services include Amazon Elastic Compute Cloud (Amazon EC2), Amazon EBS, and Amazon Virtual Private Cloud (Amazon VPC).



AWS shared responsibility model for infrastructure services

Shared responsibility model for container services

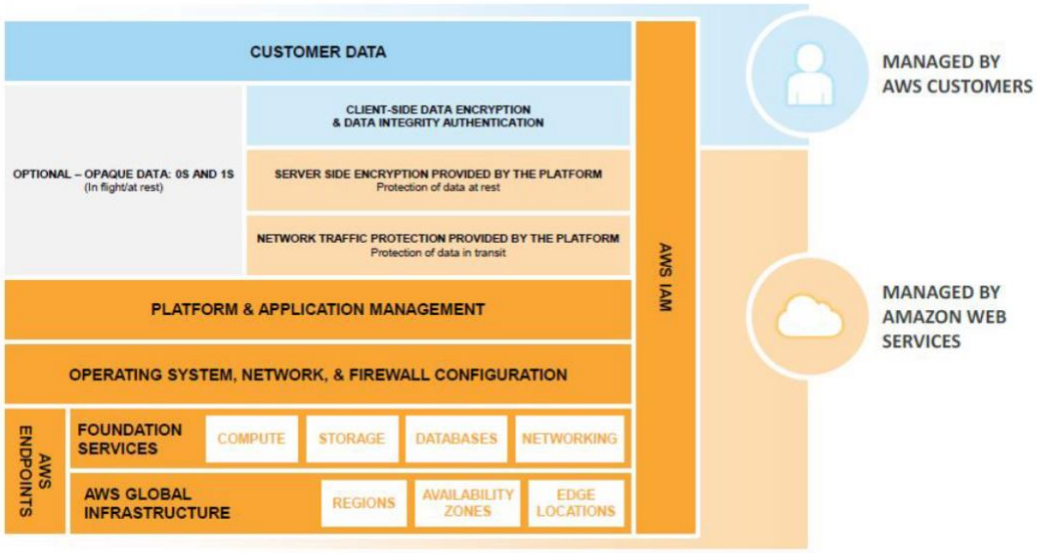
Container services include Amazon RDS and Amazon EMR.



AWS shared responsibility model for container services

Shared responsibility model for abstract services

Abstract services include Amazon S3, Amazon DynamoDB, and AWS Lambda.



AWS shared responsibility model for abstract services

The customer/AWS shared responsibility model also extends to IT controls and compliance: inherited controls, shared controls, and customer-specific controls. Just as the responsibility to operate the IT environment is shared between AWS and its

customers, so is the management, operation and verification of IT controls shared. You can add your own specific controls on top of any provided by AWS. To learn more, see the AWS Shared Responsibility Model.

AWS compliance assurance programs

AWS has obtained certifications and independent third-party attestations for a variety of industry-specific workloads.

- **SO/IEC 27001:2013** — Specifies security management best practices and comprehensive security controls following the ISO 27002 best practice guidance. The basis of this certification is the development and implementation of a rigorous security program, which defines how AWS perpetually manages security in a holistic, comprehensive manner. For more information, or to download the certification, see the AWS ISO/IEC 27001:2013 compliance page.
- **ISO/IEC 27017:2015** — Provides guidance on the information security aspects of cloud computing and recommends cloud-specific security controls that supplement the guidance of the ISO 27002 and ISO 27001 standards. This code of practice provides implementation guidance specific to cloud service providers. For more information, or to download the certification, see the AWS ISO/IEC 27017:2015 compliance page.
- **ISO/IEC 27018:2014** — Focuses on the protection of personal data in the cloud and provides implementation guidance on ISO 27002 controls applicable to personally identifiable information (PII). For more information, or to download the certification, see the AWS ISO/IEC 27018:2014 compliance page.
- **ISO 9001:2015** — Outlines a process-oriented approach to documenting and reviewing the structure, responsibilities, and procedures required to achieve effective quality management within an organization. For more information, or to download the certification, see the AWS ISO 9001:2015 compliance page.
- **PCI DSS Level 1** — The Payment Card Industry Data Security Standard (PCI DSS) is a proprietary information security standard administered by the Payment Card Industry (PCI) Security Standards Council. PCI DSS applies to all entities that store, process, or transmit cardholder data (CHD) and/or sensitive authentication data (SAD) including merchants, processors, acquirers, issuers, and service providers. The PCI DSS is mandated by the card brands and administered by the Payment Card Industry Security Standards Council. For

more information, or to request the PCI DSS Attestation of Compliance and Responsibility Summary, see the AWS PCI DSS compliance page.

SOC

AWS System and Organization Control (SOC) reports are independent third-party examination reports that demonstrate how AWS achieves key compliance controls and objectives. These reports help customers and their auditors understand the AWS controls established to support operations and compliance. For more information, see the AWS SOC compliance page. There are three types of reports:

- **SOC 1**: Provides information about the AWS control environment that may be relevant to a customer's internal controls over financial reporting, as well as information for assessing the effectiveness of internal controls over financial reporting (ICOFR).
- **SOC 2**: Provides customers, and their service users with a business need, with an independent assessment of the AWS control environment relevant to system security, availability, and confidentiality. The 2020 SOC 2 Type I Privacy report is available. The scope of the privacy report includes information about how AWS handles the content that you upload to AWS, and how it is protected in all of the services and locations that are in scope for the latest AWS SOC reports. You can download the latest SOC 2 Type I Privacy report through AWS Artifact in the AWS Management Console (sign-in required)
- **SOC 3**: Provides customers and their service users with a business need, with an independent assessment of the AWS control environment relevant to system security, availability, and confidentiality without disclosing AWS internal information.

For more information about other AWS certifications and attestations, see the AWS Compliance Programs page. For information about general AWS security controls and service-specific security, see the *AWS Overview of Security Processes* whitepaper.

AWS Artifact

AWS Artifact, an automated compliance reporting portal in the AWS Management Console, lets you review and download reports and details about more than 2,500 security controls. AWS Artifact provides on-demand access to AWS security and compliance documents. These documents include System and Organization Control (SOC) Reports, Payment Card Industry (PCI) reports, and certifications from accreditation bodies across geographies and compliance verticals.

AWS recommends the establishment of principles for data handling, security, and privacy for your customer data. These principles provide a framework for making secure and privacy enhancing decisions to achieve business outcomes. Sample principles for each category are listed below:

Data handling tenets for personal data:

- Obfuscate, de-identify, or tokenize data. Don't use the direct result when possible.
- Pass actions, not data.
- No need, no access rights. All access to data must be need-to-know to protect organizational and consumer privacy.
- Delegate data to more secure systems. Point your systems towards compliant data stores instead of sending the data itself.
- Protect this data like it's yours — because it is.
- Given the fluid regulatory environment, don't focus on the letter of any one regulation; focus on protecting your own valuable data and reducing risk for the organization and the consumer.

Security tenets for personal data:

- Keep humans away from the (raw) data.
- Compartmentalize and limit data access per business need.
- Use preventive controls over Detective controls.
- Use Detective controls to monitor preventive controls.
- Make security decisions that organization and consumers would be proud of.
- Every interaction with data should be auditable.

Privacy tenets for personal data:

- Minimize physical PII and destroy when no longer needed. PII should only be printed when absolutely necessary. Be sure to use proper disposal procedures – like shredding or placing in shred bins – and do not move or copy the data.
- Minimize data. Only accept and use what is absolutely necessary to accomplish the task.

- Expire data. Keep the data for as little time as possible via robust time-to-live (TTL) and retention policies.
- Maintain logs of all interactions with PII, including lineage.
- Keep detailed records of data subject consent, including opt-in and opt-out.

Cloud adoption framework

AWS Professional Services created the AWS Cloud Adoption Framework (AWS CAF) to help organizations successfully migrate to the cloud. AWS CAF guidance and best practices provide a comprehensive approach to cloud computing across your organization. AWS CAF helps you create an actionable, enterprise-wide cloud migration plan for your organization.

Similarly, the NIST Privacy Framework is a voluntary and customizable tool that encourages cross-organizational coordination in managing privacy risks by creating equivalence between privacy risks and other risks within your organization. The NIST Privacy Framework, used in conjunction with the AWS CAF, should make it easier for customers to move privacy practices to the cloud. The NIST Privacy Framework and the Cloud Adoption Framework, which are agnostic to law and technology, help customers manage their organization's privacy risks. See the following blog post for mapping the NIST Privacy Framework to the AWS Cloud Adoption Framework.

Preparing for the CCPA

This whitepaper discusses three major components of the CCPA:

- Data collection
- Data retrieval and deletion
- Data awareness

To help address CCPA requirements, your business can focus on these three components through the use of the following AWS services and solutions:

- **Data collection**—The following AWS services can be used to help with data collection:
  - Amazon S3
  - AWS purpose-built databases

  - AWS Lake Formation

  You can identify and manage access to personal information by using S3 object metadata, object tagging, and lifecycle management. Together, these techniques can enable you to securely collect requested personal information.

- **Data retrieval and deletion**—The following AWS services can be used to help retrieve and delete data upon request:
  - Amazon S3
  - Amazon EMR
  - AWS Glue
  - Amazon Athena
  - Amazon QuickSight

  Together, these services enable you to crawl, catalog, and query your content to retrieve specific consumer data. From there, you can further visualize the data retrieved, and use AWS CloudTrail, Amazon CloudWatch, and AWS Lambda for deletion.

- **Data awareness** — The following AWS services can be used to help notify and inform consumers about their personal information with regard to CCPA requirements:
  - Amazon Simple Email Service (Amazon SES)
  - Amazon Connect
  - Amazon Lex

  These services provide ways to notify consumers through a hosted application or by telephone.

CCPA data collection

The CCPA requires that businesses “inform consumers as to the categories of personal information to be collected and the purposes for which the categories of personal information shall be used.” Examples of personal information under the CCPA, though not an exhaustive list, include “his or her name, signature, social security number, physical characteristics or description, address, telephone number, passport number, driver’s license or state identification card number, insurance policy number, education,

employment, employment history, bank account number, credit card number, debit card number, or any other financial information, medical information, or health insurance information." Personal information could also include information such as internet browsing history, geolocation data, fingerprints, and inferences from other personal information that could create a profile about a person's preferences and characteristics.

AWS offers services and tools to help you build data ingestion architectures to categorize consumer data. By categorizing or tagging consumer data as it enters the AWS Cloud, you can separate, sort, and track personal information in your environment. Depending on the type of data and the business use case, you can store the data in various locations within AWS.

Amazon S3

Amazon S3 is a performant, secure, and feature-rich object storage service. With S3, organizations of all sizes and industries can store any amount of data for a range of use cases, including applications, Internet of Things (IoT), data lakes, analytics, backup and restore, archive, and disaster recovery. S3 is designed for 99.999999999% durability to protect data from site-level failures, errors, and threats, so it is available to your end users and applications at all times.

S3 object metadata

One method for attaching additional information to a piece of consumer data is through the use of S3 object metadata. Each Amazon S3 object has data, a key, and metadata. The object key, or key name, uniquely identifies the object in a bucket. Object metadata is a set of name-value pairs. User-defined metadata is limited to two KB in size. You can set object metadata at the time you upload the object by populating the key-value pair associated with the relevant personal information. An example would be to populate the metadata for an image of a consumer's driver's license, and so on, following your personal information's labeling standards.

Object tagging

Amazon S3 includes a feature called object tags, which, when used in combination with AWS IAM, can granularly control access to objects in Amazon S3. Object tags are user-created key-value pairs that you can add to S3 buckets or objects. You can define up to ten tags per object. S3 object tags provide two important features relevant to this use

case: IAM integration, and lifecycle management. S3 object level tags are propagated when data is moved into another AWS service.

Managing access to personal information

Within the data collection process, it's important to consider access controls for the collected data. One mechanism for controlling access is using S3 object tags and IAM policies. Object tags integrate with AWS IAM to enable your security team to control access to AWS service APIs and to specific resources. You can create IAM policies on buckets, users, or roles, and have your team test object tags for access control purposes. The following permissions policy grants a user the ability to read objects, but the condition limits that ability to only objects that have a specific tag key and value:

```
x-amz-meta-drivers-license" ="true"
x-amz-meta-consumer-name" = "true", "x-amz-meta-address"= "true"

 {
  "Version": "2012-10-17",
  "Statement": [
 {
  "Version": "2012-10-17",
  "Statement": ["Effect": "Allow",
"Action": ["s3:GetObject"
],"Resource": [ "arn:aws:s3:::examplebucket/PII-data.txt"
], "Condition": {
        "StringEquals": {  "s3:ExistingObjectTag/SSN": "true"
        }} }} }
```

This policy provides fine-grained authorization policies that limit access to personal information within the organization. As a security best practice, AWS recommends that customers follow the principle of least privilege and segregation of duties. When you create IAM policies, only grant the permissions that are required to perform a task. Determine what users need to do, and then craft policies that let them perform only those tasks.

Lifecycle management

Data collection should include designing for the entire lifecycle of the data use case. A lifecycle configuration is a set of rules that define the actions that S3 applies to a group of objects.

There are two types of actions:

- **Transaction actions** — Defines when objects transition to another storage class.
- **Expiration actions** — Defines when objects expire. Amazon S3 deletes expired objects on your behalf.

Lifecycle management in S3 typically acts as a cost optimization mechanism. However, lifecycle configuration also can be used against object tags to provide another mechanism to collect and organize personal information data. An example of this is creating a lifecycle configuration that deletes consumer objects with a “Social Security” tag after a specified period of time. Another example would be moving objects with certain custom tags to a less expensive storage class within S3, or to Amazon S3 Glacier for data archiving and long-term backup.

Automating data tagging

When writing data to S3, your application can either write associated metadata and tags directly with the API call, or attach them to the data after it’s in AWS using an abstraction layer. Both methods accomplish the goal of tagging your data in order to categorize aspects of the personal information. However, they each have their own advantages and drawbacks.

Client-side

The first option is to set the metadata and tags within the application code when you are uploading the data to S3. If you are using object tagging, S3 supports the following API operations that are specifically for object tagging:

- PutObjectTagging — Replaces tags on an object. You specify tags in the request body. There are two distinct scenarios of object tag management using this API.
- GetObjectTagging — Returns the tag set associated with an object. Returns object tags in the response body.

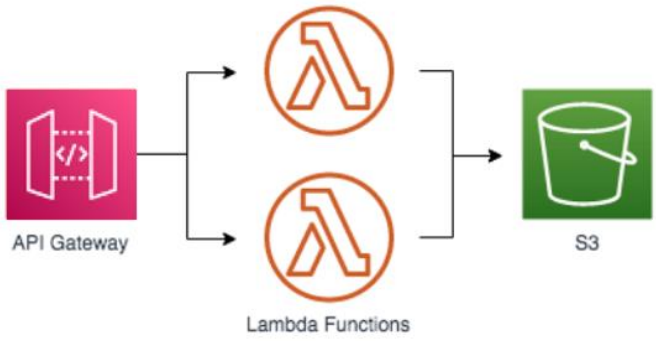
- DeleteObjectTagging — Deletes the tag set associated with an object.
  This option requires less initial setup than the server-side abstraction layer method. However, it puts the responsibility and overhead of tagging each piece of data onto the application layer. This means that developers must be aware of the tagging and metadata policy in your organization.

Server-side

The server-side approach abstracts some of the tag and metadata management away from the application itself. Instead, you reduce the development overhead by consolidating the tagging and metadata policy to an abstraction layer in between the application and Amazon S3. To design such an architecture, you can use AWS Lambda and Amazon API Gateway. AWS Lambda enables you to run code without provisioning or managing servers. Amazon API Gateway is a fully managed service that makes it easy for developers to create, publish, maintain, monitor, and secure APIs at any scale.

These services enable you to create an API backed by Lambda functions to complete the tagging of data. This frees your developers from maintaining your tagging policy. Instead, the tagging policy is set in the Lambda function and used by all uploads to S3. The Lambda function can either take inputs, such as personal information categories and customer ID, or it can tag data itself by checking the object content.

The following diagram shows an API Gateway instance backed by two Lambda functions. These functions can each implement tagging logic for different business cases of an organization's tagging policy. For example, the first Lambda function can handle the tagging of customer data with a particular sensitivity, such as social security number (SSN), customer name, and customer address. The second Lambda function can handle custom tagging from the application developer. The benefit of this approach is that the developer only needs to pass certain category flags to the API call, and the Lambda function implements the actual tagging, in accordance with the business policy.



An API Gateway instance backed by two Lambda functions

Amazon DynamoDB

Amazon DynamoDB is a key-value and document database that delivers single-digit millisecond performance at any scale. It's a fully managed, multi-Region, multi-active, durable database with built-in security, backup, restore, and in-memory caching for internet-scale applications. DynamoDB can handle more than ten trillion requests per day and supports peaks of more than 20 million requests per second.

When storing personal information in S3, you can use DynamoDB as a key-value store that holds the personal information's categories for a particular S3 object. Consider using DynamoDB as the data store for your object metadata if your per object tagging needs exceed the ten-tag limit, or if the object metadata exceeds the 2KB limit. You also can store personal information data directly in DynamoDB and use a secondary index to add the associated category to the data entry.

Amazon DynamoDB provides fast access to items in a table by specifying primary key values. However, many applications might benefit from having one or more secondary, or alternate, keys available, to allow efficient access to data with attributes other than the primary key.

A secondary index is a data structure that contains a subset of attributes from a table, along with an alternate key to support `Query` operations. You can retrieve data from the index using a `Query`, in much the same way as you use `Query` with a table. A table can have multiple secondary indexes, which gives your applications access to many different `Query` patterns.

Another key consideration is encrypting your personal information data at rest and in transit. All user data stored in Amazon DynamoDB is fully encrypted at rest. DynamoDB

encryption at rest provides enhanced security by encrypting all your data at rest using encryption keys stored in AWS Key Management Service (AWS KMS). This feature helps reduce the operational burden and complexity involved in protecting sensitive data. DynamoDB provides an additional layer of data protection by securing your data in the encrypted table, including its primary key, local and global secondary indexes, streams, global tables, backups, and DynamoDB Accelerator (DAX) clusters, whenever the data is stored in durable media.

Amazon RDS

Amazon RDS is a managed service that makes it easy to set up, operate, and scale a relational database in the cloud. It provides cost-efficient, resizable capacity and manages common database administration tasks.

One method for classifying and enhancing the protection of personal information in Amazon RDS is by using tokenization. Tokenization replaces sensitive data with unique identifiers. Some types of tokenization can also maintain the size and structure of the original data to maintain a minimum level of data utility. You can then use these identifiers to find the original sensitive data in another data source. In contrast, encryption applies a cypher to sensitive data in place so that the data is encoded in a way that only authorized parties can read it.

Tokenization is an alternative to encryption that can help to protect certain parts of the data that have high sensitivity or a specific regulatory compliance requirement. Separating the sensitive data into its own, dedicated, secured data store and using tokens in its place can help avoid the potential cost and complexity of end-to-end encryption. It can also help reduce risk through the use of temporary, one-time-use tokens.

Tokenization and encryption can be used together. You can encrypt your Amazon RDS database (DB) instances and snapshots at rest by enabling the encryption option for your Amazon RDS DB instances. Data that is encrypted at rest includes the underlying storage for DB instances, its automated backups, read replicas, and snapshots.

Amazon Redshift

Amazon Redshift is a fully managed, petabyte-scale data warehouse service in the cloud. You can start with just a few hundred gigabytes of data and scale to a petabyte or more. This enables you to acquire new insights from your data for your business and customers. When storing and reading data from Amazon Redshift, you take advantage

of the massively parallel processing (MPP) data warehouse architecture to parallelize and distribute SQL operations, to take advantage of all available resources. As a result, you can label your data with key-value pairs that are associated with the relevant personal information categories.

Enable database encryption for your clusters in Amazon Redshift to help protect your data at rest. When you enable encryption for a cluster, the data blocks and system metadata are encrypted for the cluster and its snapshots.

When considering the deletion of consumer personal information, consider the retention period of Amazon Redshift snapshots. Automated snapshots retain data until the end of the retention period. However, by default, manual snapshots taken of the Amazon Redshift cluster are retained indefinitely—even after you delete your cluster. You can change the retention period for a manual snapshot by modifying the manual snapshot settings.

Snapshot considerations

Personal information found in snapshots, whether on Amazon Redshift, Amazon RDS, or Amazon EBS, needs to be considered in your data lifecycle. If customer data is deleted after a snapshot is taken, that data can still be retrieved from the snapshot itself. There are a few methods to handle and organize this process:

- Restrict access to snapshots in accordance to the principle of least privilege and segregation of duties. This means that only individuals within your organization who need access to snapshots for critical reasons should have access to manage them.
- Delete old snapshots and create a new snapshot after customer data is deleted on production systems.
- Generate a list of customer data deletions with associated timestamps. Upon restoration of a snapshot, re-run the deletion of data between when the snapshot was taken and the current time.

These methods should be considered with compliance and business objectives in mind.

Tagging

AWS enables you to assign metadata to your AWS resources in the form of tags. Each tag is a simple label consisting of a customer-defined key and an optional value that can

make it easier to manage, search for, and filter resources by purpose, owner, environment, or other criteria. AWS tags can be used for many purposes.

AWS IAM policies support tag-based conditions, enabling customers to constrain permissions based on specific tags and their values. For example, IAM user or role permissions can include conditions to limit access to specific data types (for example, health information or financial information) or Amazon VPC networks based on their tags.

You can assign tags to identify resources that require heightened security risk management practices, for example, Amazon EC2 instances hosting applications that process sensitive or confidential data. This enables automated compliance checks to verify that proper access controls are in place, patch compliance is up to date, and so on.

Some AWS general best practices for tagging include:

- Tag everything
- Use tags consistently
- Focus on required and conditionally required tags for sensitive data
- Implement a tag governance process
- Remediate untagged resources
- Lock down tags used for access control
- Propagate tag values across related resources
- Constrain tag values with AWS Service Catalog
- Use automation to proactively tag resources
- Use compound tag values judiciously
- Integrate with authoritative data sources

Tag policies enable you to define rules on how tags can be used on AWS resources in customer accounts in AWS Organizations. Tag policies enable customers to easily adopt a standardized approach for tagging AWS resources.

With tag policies, you have a simple way to ensure your developers apply consistent tags, audit tagged resources, and maintain proper resource categorization. Standardizing tags enables customers to confidently leverage capabilities such as

attribute-based access control for critical use cases, because they are assured that resources are tagged with the right attributes.

Zero Trust

Zero Trust (ZT) security model dictates different system components to continually evaluate trust. The following is a working definition of Zero Trust, according to NIST special publication 800-207:

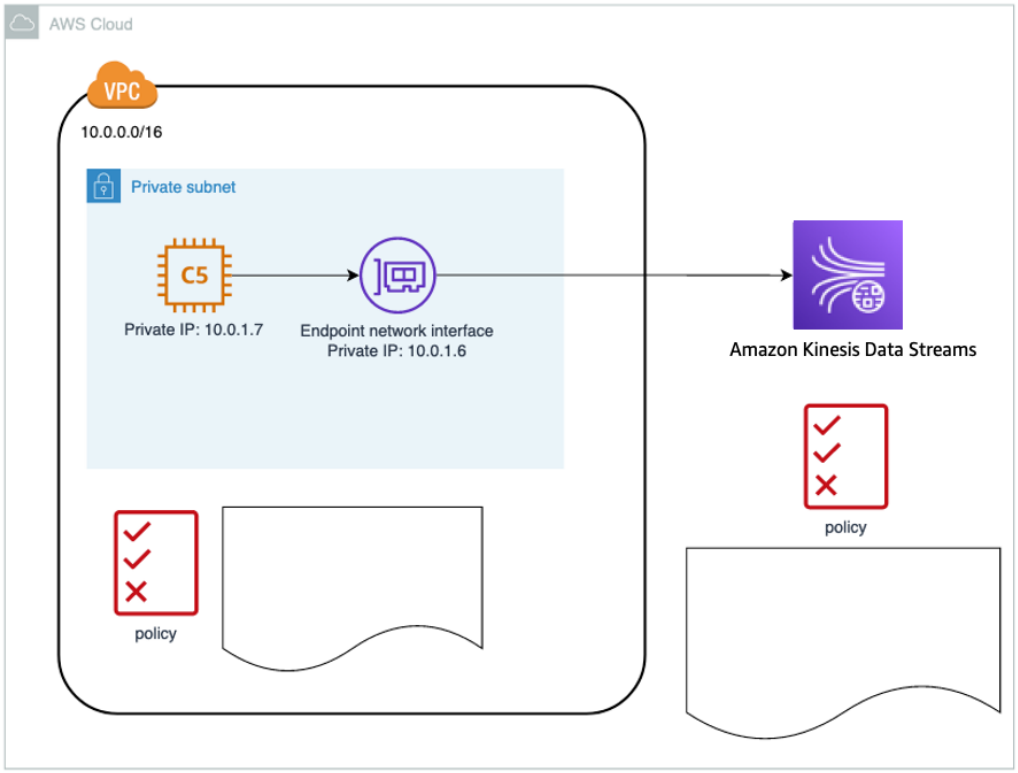
> Zero trust (ZT) provides a collection of concepts and ideas designed to minimize uncertainty in enforcing accurate, least privilege per-request access decisions in information systems and services in the face of a network viewed as compromised.
>
> — NIST special publication 800-207

The Zero Trust security model revolves around protecting data. Implementation of Zero Trust should encompass identifying sensitive data, limiting access to sensitive data and detecting any threats to sensitive data. Network and identity are two dimensions of Zero Trust.

From a networking perspective of Zero Trust, there is no inherit trust between network location or network boundaries. There is no default communication path between locations or boundaries. Instead, there are purpose-built communication paths built on micro-perimeters, microservices architecture, and packet-level controls. Cloud networking technologies for this are PrivateLink, VPC endpoints, Nitro enclaves, transit gateway, and security groups.

PrivateLink enables purpose-built network communication paths, enabling you to establish policies on what resources are allowed to traverse the network with fine grained least privileged permissions. PrivateLink is the technology that enables VPC endpoints. Using VPC endpoints, you can access AWS services hosted on the public network in a private manner.

For example, you can access an Amazon Kinesis data stream endpoint from a VPC private subnet without going over the public internet. You can configure fine-grained policies on what resources are allowed to take the network path using the VPC interface endpoint, and what API actions are permitted on the target resource. Here is a full list of AWS services you can use with AWS PrivateLink.



Purpose-built Network Communication Path Using PrivateLink VPC Interface Endpoint

There are multiple ways to implement Zero Trust identity-centric domains. For example, you can architect so that Amazon VPC serves as a trusted entity, and anything outside the VPC is not trusted. AWS Control Tower enables setting up a multi-account environment with recommended best practices out of the box. With Control Tower, you can limit access by organization ID or by principal path within the organization. This implements an architecture that contains no inherent trust between accounts within the organization. Using Service Control Policies (SCP), you can establish trust boundaries and reduce attack surface.

For a detailed discussion on Zero Trust, see Zero Trust architectures: An AWS perspective.

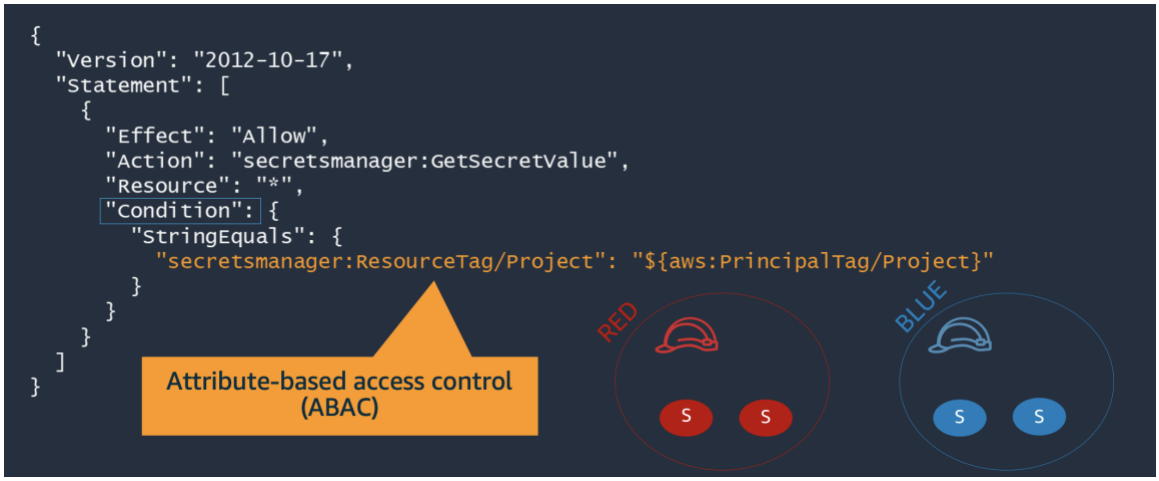
Attribute-based access control

Attribute-based access control (ABAC) is an authorization strategy that defines permissions based on attributes in conjunction with the use of policies. In AWS, these attributes are called *tags*. Tags can be attached to IAM principals (users or roles) and to AWS resources. You can create a single ABAC policy, or a small set of policies for your IAM principals. These ABAC policies can be designed to allow operations when the

principal's tag matches the resource tag. ABAC is helpful in environments that are growing rapidly, and helps with situations where policy management becomes cumbersome. ABAC provides dynamic, context-aware, and risk-intelligent access control to resources. Access control policies that include specific attributes from many different information systems can enable you to resolve complex authorization challenges and help you achieve regulatory compliance efficiently.

ABAC — IAM policy example

In this example, the action (`GetSecretValue`) is allowed only when the condition evaluates to `True`. The condition parameters contain attributes related to resource and project. This allows for reading secrets only when the project tag is matched.



Attribute-based Access Control Example

CCPA data retrieval and deletion

The CCPA also grants consumers the right to request deletion of personal information; therefore, businesses could be required to delete data upon receipt of a verified request. The following AWS services can help customers comply with this requirement by assisting in the retrieval and further deletion of specific data.

Amazon EMR

Amazon EMR is a highly distributed computing framework that enables you to quickly and easily process and store data in a cost-effective manner. Amazon EMR uses Apache Hadoop, an open-source framework, to distribute your data and processing across a resizable cluster of Amazon EC2 instances. You can also use the most

common Hadoop tools, such as Hive, Pig, and Spark. Hadoop provides a framework to run big data processing and analytics. Amazon EMR does all the work involved with provisioning, managing, and maintaining the infrastructure and software of a Hadoop cluster.

AWS Glue

AWS Glue is a fully managed extract, transform, and load (ETL) service that you can use to catalog your data, clean it, enrich it, and move it reliably between data stores. With AWS Glue, you can significantly reduce the cost, complexity, and time spent creating ETL jobs. AWS Glue is serverless, so there is no infrastructure to set up or manage. You pay only for the resources consumed when your jobs are running.

AWS Glue connects to the data source of your choice, such as an S3 file, an Amazon RDS table, Amazon EMR, or another set of data being used for collection. As a result, all of your data is stored and available as it pertains to that data store's durability characteristics. When necessary for CCPA purposes, you can run a serverless Apache Spark job within AWS Glue on your data store to retrieve specified data or to remove personal information data.

Using AWS Glue gives you the following benefits, which may help with the retrieval and deletion of data:

- AWS Glue can automatically crawl your data and generate code for ETL processes. For example, you could write an AWS Glue job to retrieve all the data within an S3 bucket for a specific user, transform that data, and prepare that data for deletion when requested.
- Integration with services like Amazon Athena, Amazon EMR, and Amazon Redshift is provided. The next section provides an example of using AWS Glue to query an S3 data lake for data retrieval using these services.
- AWS Glue is serverless — there is no infrastructure to provision or manage. A managed ETL service is provided that runs in a serverless Apache Spark environment. This capability allows you to focus on CCPA compliance in your ETL job, and not have to worry about configuring and managing the underlying compute resources.
- AWS Glue generates ETL code that is customizable, reusable, and portable, using familiar technology — Python and Spark. After you've customized or written a job for a particular CCPA task, you can reuse that script for additional consumer requests.

AWS Glue DataBrew

AWS Glue DataBrew is a no-code data preparation service that enables data analysts and data scientists to prepare data via an interactive, point-and-click visual interface, reducing the time it takes to prepare data by up to 80%. It empowers data analysts and data scientists to easily understand, clean, and transform data. Users can also explore and understand their data by looking at visual representations of how their data is distributed.

DataBrew provides recommendations to resolve data quality issues such as incorrectly classified data types and missing values, and also includes over 250 built-in transformations. This makes it easy for users to combine data from multiple sources and perform multiple data quality and data transformations tasks such as removing nulls, fixing schema inconsistencies, adding data pivots, or merging and splitting columns into a single, repeatable workflow known as a recipe.

DataBrew can help protect personal data while maintaining data utility tokenization. For more complex transformations such as automatically replacing missing values, or converting words to a common base or root word, Elixir uses advanced machine learning algorithms and techniques such as Natural Language Processing (NLP). Data engineers can then integrate these recipes in production pipelines using APIs and schedule them to run on a recurring basis to keep their dataset up to date.

Once prepared, Elixir pushes the data to an S3 data lake. With the prepared data now available in S3, customers can also benefit from analyzing that data using a variety of analytics services such as Amazon Redshift, Amazon Athena, and Amazon SageMaker, as well as BI tools such as Amazon QuickSight or Tableau. DataBrew can also be used to implement a key aspect of CCPA: the ability to track and map data lineage. Mapping data lineage assists in tracking the various data sources and transformation steps that the data has been through.

Amazon Athena

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to set up or manage, and you can start analyzing data immediately. You don’t need to load your data into Athena, as it works directly with data stored in S3. Just log in to the Athena Console, define your table schema, and start querying. Amazon Athena uses Presto with full ANSI SQL support, and works with a variety of standard data formats, including CSV, JSON, ORC, Apache Parquet, and Apache Avro.

Amazon QuickSight

Amazon QuickSight is a fast, cloud-based business analytics service that makes it easier for employees within an organization to build visualizations, perform ad hoc analysis, and quickly get business insights from their data, anytime, on any device. It can connect to a wide variety of data sources including flat files such as CSV and Excel, on-premises databases such as SQL Server, MySQL, and PostgreSQL, and AWS resources such as Amazon RDS databases, Amazon Redshift, Amazon Athena, and Amazon S3. Amazon QuickSight enables organizations to scale their business analytics capabilities to hundreds of thousands of users, and delivers fast and responsive query performance.

Amazon QuickSight is built with a Super-fast, Parallel, In-memory Calculation Engine (SPICE). Built for the Cloud, SPICE uses a combination of columnar storage and in-memory technologies (enabled through the latest hardware innovations and machine code generation) to run interactive queries on large datasets and get rapid responses. SPICE supports rich calculations to help you derive valuable insights from your analysis without needing to provision or manage infrastructure. Data in SPICE persists until it's explicitly deleted. SPICE also automatically replicates data for high availability and enables Amazon QuickSight to scale to hundreds of thousands of users who can all simultaneously perform fast interactive analysis across a wide variety of AWS data sources.

Amazon CodeGuru Reviewer

Amazon CodeGuru Reviewer is a service that uses program analysis and machine learning to detect potential defects that are difficult for developers to find and offers suggestions for improving your Java and Python code.

By proactively detecting code defects, CodeGuru Reviewer can provide guidelines for addressing them and implementing best practices to improve the overall quality and maintainability of your code base during the code review stage. CodeGuru Reviewer doesn't flag syntactical mistakes, as these are relatively easy to find. Instead, CodeGuru Reviewer will identify more complex problems and suggest improvements related to the AWS best practices, resource leak prevention, and sensitive information leak prevention, among others. Amazon CodeGuru currently supports Java and Python.

Queries against an Amazon S3 data lake for data retrieval

There are several options for storage on AWS. Data lakes are an increasingly popular way to store and analyze both structured and unstructured data. With a single source of aggregation for all of your data, a data lake is a useful starting point for data retrieval and later deletion. If you use an S3 data lake, AWS Glue can make all your data immediately available for analytics without moving the data.

AWS Glue crawlers scan your data lake and keep the AWS Glue Data Catalog in sync with the underlying data. You can then directly query your data lake with Amazon Athena and Amazon Redshift Spectrum. In particular, this architecture enables you to, for example, query directly for a unique customer ID, collect the entirety of their data from your data lake, and promptly delete it. You also can use the AWS Glue Data Catalog as your external Apache Hive Metastore for big data applications running on Amazon EMR. The following list outlines an architecture pattern to query your S3 data lake for data retrieval:

- An AWS Glue crawler connects to a data store, progresses through a prioritized list of classifiers to extract the schema of your data and other statistics, and then populates the AWS Glue Data Catalog with this metadata. Crawlers can run periodically to detect the availability of new data and changes to existing data, including table definition changes.

  As new consumer data is added to your data lake, AWS Glue crawlers will regularly update the Glue Data Catalog accordingly. Crawlers automatically add new tables, new partitions to existing tables, and new versions of table definitions. You can customize AWS Glue crawlers to classify your own file types.

- The AWS Glue Data Catalog is a central repository to store structural and operational metadata for all your data assets. For a given dataset, you can store its table definition, physical location, add business-relevant attributes, and track how this data has changed over time.

  The AWS Glue Data Catalog is Apache Hive Metastore-compatible and is a drop-in replacement for the Apache Hive Metastore for Big Data applications running on Amazon EMR. For more information, see Using the AWS Glue Data Catalog as the metastore for Hive.

- The AWS Glue Data Catalog also provides out-of-the-box integration with Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum. After you add your table definitions to the AWS Glue Data Catalog, they are available for ETL and also readily available for querying in Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum. This enables you to have a common view of your data between these services.

  Amazon Athena enables you to view and query data within your data lake on a consumer level without having to move your data from the data lake. This enables you to easily retrieve data upon consumer request from your primary data store in S3.

- Using a business intelligence (BI) tool such as Amazon QuickSight enables you to easily build visualizations, perform ad hoc analysis, and quickly get business insights from your data. Amazon QuickSight supports data sources such as Amazon Athena, Amazon Redshift Spectrum, Amazon S3, and many others. Through the use of Amazon QuickSight, you can further visualize the data retrieved for deletion.

Additionally, if you build your S3 data lake using AWS Lake Formation, you can use the Data Catalog search capabilities to easily search across databases and tables within your data lake. This feature enables you to query for specific properties such as the customer ID of a consumer and to identify all metadata across the data lake containing this specific value.

In addition, Lake Formation enables searches across your entire data lake by keyword, as well as the ability to apply multiple filters at once. If a consumer can be identified in more than one way (for example, a user ID, group ID, and so on), using these search capabilities helps ensure that all data associated with a certain consumer has been retrieved.

Opt-out of personal information sales requests

The CCPA also grants “a consumer ... the right to request that a business that sells the consumer’s personal information, or discloses it for a business purpose, disclose” the categories of information that it collects and categories of information and the identity of third parties to which the information was sold or disclosed. The CCPA requires “businesses” that sell personal information to provide this information in response to a verifiable consumer request.

Several of the AWS services previously mentioned can help you comply with this requirement. Amazon QuickSight allows end users to create visualizations and provide

insights into their data. In particular, through the use of QuickSight dashboards, you can create, sort, and filter the retrieved data by category and then provide this information to your consumers. Using this information, consumers may choose to opt out of the sale of their personal information.

Monitoring and logging for further data deletion

Monitoring and logging of your AWS environment is a critical component of IT governance, security, and compliance. AWS CloudTrail provides a simple solution to record AWS API calls and resource changes. CloudTrail helps alleviate the burden of on premises infrastructure and storage challenges by helping you to build preventative and detective security controls.

On premises logging solutions require installing agents, setting up configuration files, using centralized log servers, and building and maintaining expensive, highly durable data stores to store the data. AWS CloudTrail eliminates this burdensome infrastructure setup and enables you to turn on logging in as few as two clicks, and get increased visibility into all API calls in your AWS account.

CloudTrail continuously captures API calls from multiple servers into a highly available processing pipeline. To turn on CloudTrail, sign into the AWS Management Console, navigate to the CloudTrail console, and click to enable logging. To learn more about services and Regions available for use, see the CloudTrail Supported Regions page. It's important to understand changes that are made to your resources, as this may impact the data collected and potentially distributed to your consumers. To learn more about logging and monitoring best practices, see the *Security at Scale: Logging in AWS* whitepaper.

After data has been collected properly, it may be necessary to delete data that is specific to a consumer. AWS Lambda and Amazon CloudWatch may help you comply with this requirement. AWS Lambda is an event-driven, serverless compute service that extends other AWS services with custom logic, or creates other backend services that operate with scale, performance, and security. Amazon CloudWatch is a monitoring and management service that can provide actionable insights about your data.

AWS Lambda can automatically run code in response to multiple events, such as HTTP requests through Amazon API Gateway, modifications to objects in S3 buckets, table updates in DynamoDB, and state transitions in AWS Step Functions. You also can run code directly from any web or mobile app. Lambda runs code on a highly available compute infrastructure, and performs all of the administration of the underlying platform,

including server and operating system maintenance, capacity provisioning, automatic scaling, patching, code monitoring, and logging.

With Lambda, you can just upload your code and configure when to invoke it. Lambda takes care of everything required to run and scale your code with high availability. You can integrate it with many other AWS services, such as Amazon CloudWatch, and create serverless applications or backend services, ranging from periodically triggered, simple automation tasks to full-fledged microservices applications.

Amazon CloudWatch can collect metrics across the resources in your architecture. You also can collect and publish custom metrics to surface specific business or other derived metrics. For example, upon receiving a deletion request by a consumer, you can use a CloudWatch Event to run an AWS Lambda function. This Lambda function may contain code to gather and delete all metadata associated with a given identifier provided by the consumer.

You can also create a verification process to confirm that the requested data was successfully deleted and the request recorded. For example, suppose that you want to delete all the data for a specified user. By creating a setup similar to the one described above, your Lambda function can identify the associated metadata for the specific user ID within an Amazon RDS table, delete the data, and record the successful deletion in a DynamoDB table.

CCPA data awareness

Knowledge and notification

Customer notification and information delivery can be performed programmatically using several mechanisms to provide secure access to customer-specific information based on their explicit request for this information. There are a variety of solutions to this customer experience component. A web-based internet application can be used to obtain customer content from a data source and securely return that information (using an HTTPS response) to the customer's browser. One example solution for hosting this application is the Reference Architecture for a Web Application on AWS.

Alternatively, you can use this same process to send an email to the customer with the relevant information using Amazon SES. Or you can establish a telephone-based system using Amazon Connect to notify the customer with a text message or voice call. Both Amazon SES and Amazon Connect can be configured to access a wide variety of

data sources, extract relevant data, and return that data to the customer using the best method for the customer and the particular situation.

Amazon SES

Amazon SES is a cloud-based email sending service designed to help digital marketers and application developers send marketing, notification, and transactional emails. It is a reliable, cost-effective service for businesses of all sizes that use email to keep in contact with their customers.

You can use the SMTP interface, or one of the AWS SDKs, to integrate Amazon SES into your existing applications. You also can integrate the email sending capabilities of Amazon SES into the software that you currently use, such as ticketing systems and email clients. Amazon SES supports standard authentication mechanisms, including DomainKeys Identified Mail (DKIM), Sender Policy Framework (SPF), and Domain-based Message Authentication, Reporting, and Conformance (DMARC).

Amazon Connect

Amazon Connect is a self-service, cloud-based contact center service that is easy to integrate with other systems such as customer relationship management (CRM) solutions or other AWS services. For example, you can use AWS Lambda to run code in a serverless application or backend service and build contact flow experiences that adapt to your customer's needs in real-time. Amazon Connect provides out-of-the-box integrations with many popular tools such as customer relationship management (CRM), workforce management (WFM), and various analytics platforms.

You also can use Amazon Connect with other AWS services, such as Amazon S3 and AWS Lambda, to store recorded calls or to stream detailed contact records in real-time to a data warehouse. Business intelligence systems can then access this data and perform further analysis. Amazon Connect provides an API so that you can customize the solution to your needs.

The contact flows in Amazon Connect can be used to create dynamic interactive voice response (IVR) solutions. With Amazon Connect, you can gather appropriate personal information to customize your customer's experience when they interact with your IVR. The personal information used for this customization can include, for example, social security numbers, credit card information, and addresses. AWS recommends that you always encrypt personal information while in motion and when stored.

Amazon Connect, when paired with Amazon Lex bots, can capture customer input entered on their numeric keypad as digits in a contact flow. Amazon Connect matches the intent based on that input in the same way that it matches the intent when you speak an utterance.

This option provides the capability to accept input in a manner that best suites each customer. Whether they choose to use touch-tone data entry or their voice, Amazon Connect can capture the input both ways and provide the same functionality. Whether you have an existing call center implementation or not, Amazon Connect can help you meet the requirements for having a toll-free number and can be configured in a way that relies more on code and automation rather than human call center agents.

You can use the store customer input block in Amazon Connect to gather sensitive personal information, as well as automatically encrypt that data using your encryption keys. This feature can help you comply with the required encryption requirements for CCPA. Amazon Connect uses the AWS Encryption SDK to encrypt data, and the SDK uses an envelope encryption approach. This approach is designed to protect both the raw data and the data keys used to encrypt them. For more information about how the AWS Encryption SDK works, see Envelope Encryption.

You can use AI services from AWS with Amazon Connect to help your organization operate more efficiently and improve the customer experience. For example, you can integrate Amazon Lex intelligent conversational bots into contact flows to turn automated interactions into natural conversations.

A customer dials a company's toll-free number and their phone number is detected by Amazon Connect, which in turn issues a query or API call to the CRM system using an AWS Lambda function. When the Lambda function is invoked, it builds a request that contains contact data, user attributes, and parameters that are specific to the Lambda function. The Lambda function is also configured to parse the event and return a simple string map. This string map is simple because it's a set of key-value pairs, it cannot contain nested attributes, and it must contain less than 32 KB of UTF-8 data.

There are two ways to use the function response in your contact flow. You can either directly reference the variables returned from Lambda, or store the values returned as contact attributes and then reference the stored attributes. When you use an external reference to a response from a Lambda function, the reference will always receive the response from the most recently invoked function. To use the response from a Lambda function in a subsequent function, the response must either be saved as a contact attribute, or passed as a parameter to the next function. If you store responses as

contact attributes, you can use them throughout your contact flow, and they are included in contact trace records (CTR).

Sample architectures

These sample architectures are meant for guidance and spurring innovation. Each architecture will need to be customized for your individual requirements. The architectures represent a range of complexity, and address various aspects of privacy compliance.

S3 Find and Forget

Amazon S3 Find and Forget is a solution that addresses the need to selectively erase records from data lakes stored on Amazon S3. This solution can assist data lake operators to handle data erasure requests; for example, pursuant to the CCPA and CPRA.

The Amazon S3 Find and Forget solution can be used with Parquet or JSON format data stored in S3 buckets. Your data lake is connected to the solution via an AWS Glue table, and specifies which columns in the table contain user identifiers.

This architecture addresses the CCPA data subject access right to deletion. Under the expansion of CCPA by CPRA, data subjects have the right to request all their personal information be deleted. Below are the key elements of this solution.

Design principles

The goal of the solution is to provide a secure, reliable, performant and cost-effective tool for finding and removing individual records within objects stored in S3 buckets. To achieve these goals the solution has adopted the following design principles:

1. **Secure by design** —
    - Every component is implemented with least privilege access
    - Encryption is performed at all layers at rest and in transit
    - Authentication is provided out of the box
    - Expiration of logs is configurable
    - Record identifiers (known as Match IDs) are automatically obfuscated or irreversibly deleted as soon as possible when persisting state

2. **Built to scale** — The system is designed and tested to work with petabyte-scale data lakes containing thousands of partitions and hundreds of thousands of objects.
3. **Cost optimized** —
    - **Perform work in batches** — Because the time complexity of removing a single record vs. multiple records in a single object is practically equal and it is common for data owners to have the requirement of removing data within a given *timeframe*, the solution is designed to allow the solution operator to "queue" multiple matches to be removed in a single job.
    - **Fail fast** — A deletion job takes place in two distinct phases: Find and Forget. The Find phase queries the objects in your S3 data lakes to find any objects which contain records where a specified column contains at least one of the Match IDs in the deletion queue. If any queries fail, the job will abandon as soon as possible and the Forget phase will not take place. The Forget Phase takes the list of objects returned from the Find phase, and deletes only the relevant rows in those objects.
    - **Optimized for Parquet** — The split phase approach optimizes scanning for columnar dense formats such as Parquet. The Find phase only retrieves and processes the data for relevant columns when determining which S3 objects need to be processed in the Forget phase. This approach can have significant cost savings when operating on large data lakes with sparse matches.
    - **Serverless** — Where possible, the solution only uses Serverless components to avoid costs for idle resources. All the components for Web UI, API and deletion jobs are serverless.
4. **Robust monitoring and logging** — When performing deletion jobs, information is provided in real-time to provide visibility. After the job completes, detailed reports are available documenting all the actions performed to individual S3 Objects, and detailed error traces in case of failures to facilitate troubleshooting processes and identify remediation actions.

Core components

The following terms are used to identify core components within the solution.

Data Mappers

Data Mappers instruct the Amazon S3 Find and Forget solution how and where to search for items to be deleted.

To find data, a Data Mapper uses:

- A table in a supported data catalog provider which describes the location and structure of the data you want to connect to the solution. Currently, AWS Glue is the only supported data catalog provider.
- A *function*, which is the service the Amazon S3 Find and Forget solution uses to query the data. Currently, Amazon Athena is the only supported query function.

Data Mappers can be created at any time, and removed when no deletion job is running.

Deletion queue

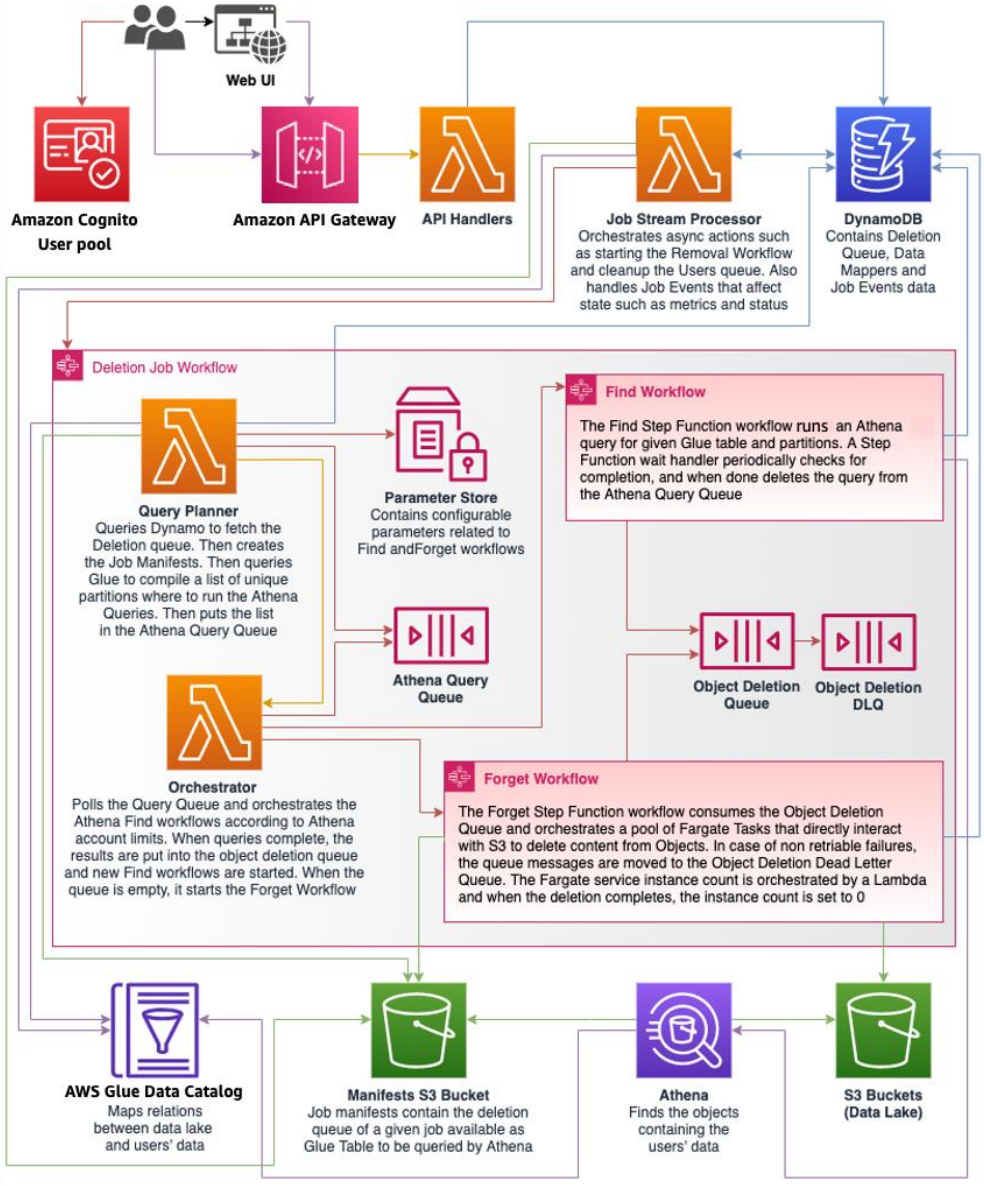
The deletion queue is a list of matches. A *match* is a value you want to search for, which identifies rows in your S3 data lake to be deleted. For example, a match could be the ID of a specific customer.

Matches can be added at any time, and can be removed only when no deletion job is in progress.

Deletion jobs

A deletion job is an activity performed by Amazon S3 Find and Forget which queries your data in S3 defined by the Data Mappers and deletes rows containing any match present in the Deletion Queue.

Deletion jobs can be run anytime there is not another deletion job already running.



Sample Privacy Architecture Showcasing Right to Erasure

Privacy bot

This architecture has data being ingested through AWS Lake Formation. Lake Formation has built in security and policy management that can be leveraged to support CCPA requirements. AWS Glue DataBrew is used for normalization, tokenization, and conformance packs. AWS Config rules are used to detect privacy violations, and CloudWatch offers real-time alerting.

After data is ingested, it is classified and scanned by Amazon Macie. Privacy bots monitor for privacy and security violations, and trigger the respective Lambda functions for audit and remediation.

For specific types of data such as healthcare, you can use Amazon HealthLake for greater assistance with HIPAA and HITRUST compliance. For data analytics, AWS AI/ML technologies such as Amazon Comprehend and Amazon CodeGuru continuously check for privacy leakage while expanding to analytics to native platforms such as Amazon EMR, Amazon Redshift, and AWS IoT.

Personal information ingestion pipeline

This is a centralized ingestion pipeline where data can be classified and metadata is collected. Collected metadata can be used later for policy enforcement actions and detailed visualization of all personal information in the environment. In this example, data is extracted from an RDS database, and normalized and transformed with AWS Glue. The resultant Parquet file is stored in S3.

Presence of the file in S3 triggers Amazon Macie to analyze the file. When Macie finds personal data, it triggers a Lambda function via SNS to capture the metadata and places it in an AWS Glue table. Lambda takes the metadata and proximity data from Macie to calculate privacy risk values, and stores them in the DynamoDB table. QuickSight is used to visualize the privacy risk of the data over time. Open Policy Agent can be used to create policy-based privacy controls at the individual data element and data aggregate levels.

Conclusion

AWS offers a wealth of services that can assist you with your CCPA data collection, data retrieval and deletion, and data awareness requirements. These services are provided in a secure and extensible manner that scales across multiple Regions and geographies. AWS services can be configured and customized to help meet the requirements of regulators and consumers. A variety of mechanisms are available for securely providing consumers with access to their personal information, including websites, email delivery, and agentless call centers.

This whitepaper also described the AWS shared responsibility model, which delineates the responsibilities between you and AWS. AWS owns security “of” the cloud, and you, the customer, owns security “in” the cloud. AWS has obtained certifications and attestations for a variety of compliance and security controls, and makes it easy to

report on more than 2,500 security controls using AWS Artifact—an automated compliance-reporting service. As you prepare for the CCPA, you may want to visit Tools to Build on AWS to learn about options for building anything from small scripts that delete data to a full orchestration framework that uses AWS Code services.

Contributors

Contributors to this document include:

- Julia Soscia, Manager, Startup Solutions Architect, Amazon Web Services
- Anthony Pasquariello, Solutions Architect, Amazon Web Services
- Justin De Castri, Solutions Architect Manager, Amazon Web Services
- Jodi Scrofani, AWS Security Global Manager, Amazon Web Services
- Marta Taggart, Senior Program Manager, Amazon Web Services
- Carl Mathis, Senior Privacy Assurance Consultant, Amazon Web Services
- Faisal Farooq, Startup Solutions Architect, Amazon Web Services

Further reading

For additional information, see:

- *AWS Best Practices for DDoS Resiliency*
- AWS Security Checklist
- Encryption of Data at Rest
- Cloud Adoption Framework - Security Perspective
- Data Classification: Secure Cloud Adoption
- AWS Security Best Practices
- Encryption of Data at Rest
- *Amazon Web Services: Risk and Compliance*
- Using AWS in the Context of Common Privacy and Data Protection Considerations

- Security at Scale: Logging in AWS
- AWS Governance at Scale
- Secure Content Delivery with Amazon CloudFront

Document versions

Date	Description
November 3, 2021	Updated for technical accuracy
July 2019	First publication