Preparing for the California Consumer Privacy Act

July 2019

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.

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

Maintaining customer trust is an ongoing commitment at AWS. We strive to inform you of our privacy and data security policies, practices, and associated technologies that we’ve put in place. 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. We provide 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. We provide APIs that allow you to control access to any of the services you develop or deploy in AWS. We do not access or use your content for any purpose without your prior consent. We never use your content, or derive 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. We will not move or replicate your content outside of your chosen AWS Regions 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. We offer you strong encryption for your content in transit and at rest, and we provide 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), and Amazon Redshift.
  - Flexible key management options, including AWS Key Management Service (KMS), which allow 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 S3-Managed Keys (SSE-S3), AWS KMS-Managed Keys (SSE-KMS), or customer-provided encryption keys (SSE-C).

- **Disclosure of customer content**: We do not disclose customer information unless we're required to do so to comply with a legally valid and binding order. Amazon notifies you before disclosing your content information, unless we are legally prohibited from doing so, or there is a clear indication of illegal conduct regarding the use of Amazon products or services.
- **Security Assurance**: We have a security assurance program that follows best practices for global privacy and data protection. These controls permit you to operate securely within AWS, and make the best use of our security control environment. These security protections and control processes are independently validated by multiple third-party independent assessments.[1]

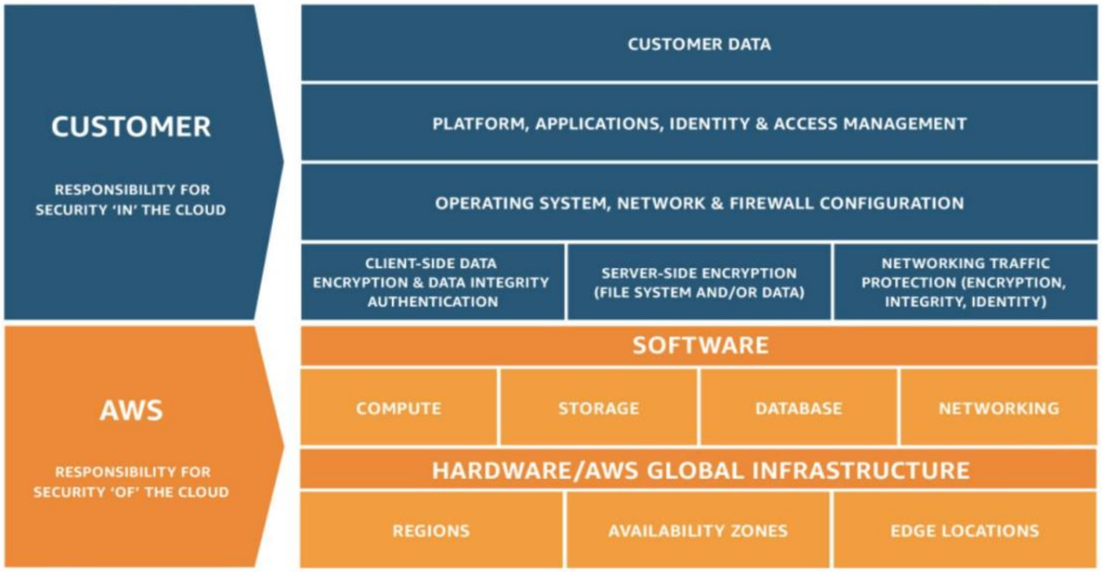
For guidance and best practices on building security policies and processes for your organization, see the AWS Security Best Practices whitepaper.[2] For information on how AWS collects and uses personal information, see our Privacy Notice.[3]

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.

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. To learn more, see the AWS Shared Responsibility Model.[4]


CUSTOMER
RESPONSIBILITY FOR SECURITY 'IN' THE CLOUD
AWS
RESPONSIBILITY FOR SECURITY 'OF' THE CLOUD
CUSTOMER DATA
PLATFORM, APPLICATIONS, IDENTITY & ACCESS MANAGEMENT
OPERATING SYSTEM, NETWORK & FIREWALL CONFIGURATION
CLIENT-SIDE DATA ENCRYPTION & DATA INTEGRITY AUTHENTICATION
SERVER-SIDE ENCRYPTION (FILE SYSTEM AND/OR DATA)
NETWORKING TRAFFIC PROTECTION (ENCRYPTION, INTEGRITY, IDENTITY)
SOFTWARE
COMPUTE
STORAGE
DATABASE
NETWORKING
HARDWARE/AWS GLOBAL INFRASTRUCTURE
REGIONS
AVAILABILITY ZONES
EDGE LOCATIONS

Shared Responsibility Model

AWS Compliance Assurance Programs

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

- **ISO/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.[5]

- **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.[6]
- **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.[7]
- **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.[8]
- **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.[9]

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.[10] 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.
- **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.[11] For information about general AWS security controls and service-specific security, see the AWS Overview of Security Processes whitepaper.[12]

AWS Artifact

AWS Artifact,[13] 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.

Preparing for the CCPA

This whitepaper discusses three major components of the CCPA: Data Collection, Data Retrieval and Deletion, and Data Awareness. In order to address CCPA requirements, your business can focus on these three components through the use of the following AWS services and solutions:

1. **Data Collection**—The following AWS services can be used to help with data collection: Amazon Simple Storage Service (Amazon S3), Amazon DynamoDB, and Amazon Redshift. You can identify and manage access to personal information by using S3 object metadata, object tagging, and lifecycle management. Together, these techniques allow you to securely collect requested personally identifiable information (PII).

2. **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, and Amazon QuickSight. Together, these services allow 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, AWS Lambda, and AWS Config for deletion.
3. **Data Awareness**—The following AWS services can be used to help notify and inform consumers about their personal information with regard to CCPA requirements: AWS Config, Amazon Simple Email Service (SES), Amazon Connect, and 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."

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 Simple Storage Service (Amazon S3[14]) is a performant, secure, and feature-rich object storage service. With Amazon S3, organizations of all sizes and industries can store any amount of data for any use case, including applications, IoT, data lakes, analytics, backup and restore, archive, and disaster recovery. Amazon S3 is designed for 99.999999999% durability to protect data from site-level failures, errors, and threats, so that 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 2 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 PII data. An example would be to populate the metadata for an image of a consumer's driver’s license with “`x-amz-meta-drivers-license" = "true"`,“`x-amz-meta-consumer-name" = "true", ‟x-amz-meta-address" = "true"`, and so on, following your PII 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 10 tags per object. S3 object tags provide two important features relevant to this use case: IAM integration, and Lifecycle Management.

Managing Access to PII Data

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 them 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:

```
{
  "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 provides internal security measures to enable fine grain authorization policies to limit access of personal data within the organization. As a security best practice, AWS recommends that customers follow the *principle of least privilege.* 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 Amazon 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 Amazon 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 PII 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 Amazon S3, or to Amazon S3 Glacier for data archiving and long-term backup.

Automating Data Tagging

When writing data to Amazon S3, your application can either write associated metadata and tags directly with the API call, or attach them to the data after it is in AWS using an abstraction layer. Both methods accomplish the goal of tagging your data in order to categorize the PII aspects. 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 Amazon S3. If you are using object tagging, Amazon S3 supports the following API operations that are specifically for object tagging:

- PUT Object tagging[15]–Replaces tags on an object. You specify tags in the request body. There are two distinct scenarios of object tag management using this API.
- GET Object tagging[16]–Returns the tag set associated with an object. Returns object tags in the response body.
- DELETE Object tagging[17]–Deletes the tag set associated with an object.

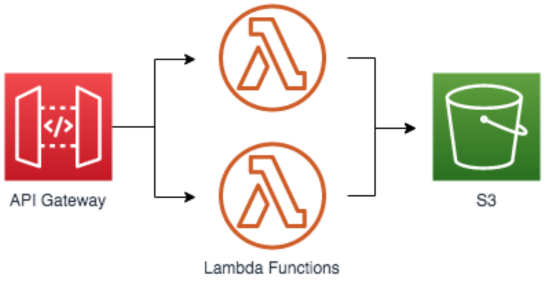
This option requires less initial setup than the server-side abstraction layer method. However, it puts the responsibility and overhead of tagging each piece of data onto the application layer. This means that developers must be aware of the tagging and metadata policy in your organization.

Server-Side Abstraction Layer

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[18] and Amazon API Gateway.[19] AWS Lambda lets you 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 allow 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 Amazon S3. The Lambda function can either take inputs, such as PII categories and customer ID, or it can tag data itself by checking the object content.

The following diagram shows an API Gateway 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 PII category flags to the API call, and the Lambda function implements the actual tagging, in accordance with the business policy.



Amazon DynamoDB

Amazon DynamoDB[20] is a key-value and document database that delivers single-digit millisecond performance at any scale. It's a fully managed, multi-region, multi-master database with built-in security, backup, restore, and in-memory caching for internet-scale applications. DynamoDB can handle more than 10 trillion requests per day and support peaks of more than 20 million requests per second.

When storing PII data into Amazon S3, you can use DynamoDB as a key-value store that holds the PII 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 10 tag limit, or if the object metadata exceeds the 2-KB limit. You also can store PII data directly into DynamoDB and use a secondary index to add the associated PII 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 PII data at rest and in transit. All user data stored in Amazon DynamoDB is encrypted at rest. DynamoDB provides enhanced security by encrypting all your data at rest using encryption keys stored in AWS Key Management Service (AWS KMS)[21]. 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 Relational Database Service (Amazon RDS[22]) 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 PII data in Amazon RDS is by using tokenization. Tokenization replaces sensitive data with unique identifiers. You 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 has high sensitivity, or a specific regulatory compliance requirement, such as PCI. 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 also can 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 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[23] 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 PII categories.

Enable database encryption for your clusters in Amazon Redshift to protect 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 PII data, make sure to 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

PII data found in snapshots, whether on Amazon Redshift, Amazon RDS, or Amazon EBS, need 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.

First, restrict access to snapshots in accordance to the principle of least privilege. This means that only individuals within your organization who need access to snapshots for critical reasons should have access to manage them. The second method to handle PII data on snapshots would be to delete old snapshots and create a new snapshot after customer data is deleted on production systems. A third method would be to generate a list of customer data deletions with associated timestamps. Upon restoration of a snapshot, you would 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.

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 help customers comply with this requirement by assisting in the retrieval and further deletion of specific data.

Amazon EMR

Amazon EMR[24] is a highly distributed computing framework that allows 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 also can 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[25] 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 Amazon S3 file, an Amazon RDS table, 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 automatically can 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.

Amazon Athena

Amazon Athena[26] 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 Amazon S3. Just log into 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

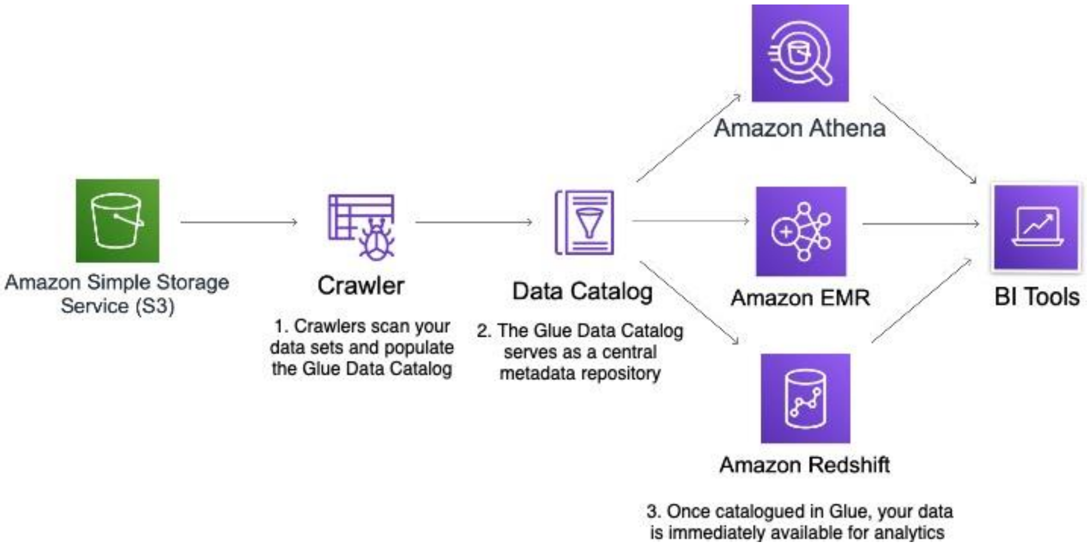
Amazon QuickSight[27] is a fast, easy-to-use, cloud-based business analytics service that makes it easy for all employees within an organization to build visualizations, perform ad hoc analysis, and quickly get business insights from their data, anytime, on any device. It can connect to a wide variety of data sources including flat files, such as CSV and Excel, access to on-premises databases, such as SQL Server, MySQL, and PostgreSQL, and AWS resources, like 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 *SPICE* – a Super-fast, Parallel, In-memory Calculation Engine. 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.

Queries Against an Amazon S3 Data Lake for Data Retrieval

There are several options for storage on AWS. However, 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 Amazon 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 allows you to query directly for a unique customer ID, for example, 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.


Amazon Simple Storage Service (S3)
Crawler
1. Crawlers scan your data sets and populate the Glue Data Catalog
Data Catalog
2. The Glue Data Catalog serves as a central metadata repository
Amazon Athena
Amazon EMR
Amazon Redshift
3. Once catalogued in Glue, your data is immediately available for analytics
BI Tools

1. 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.

2. 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.[28]

3. The AWS Glue Data Catalog also provides out-of-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 allows 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 allows you to easily retrieve data upon consumer request from your primary data store in Amazon S3.

4. Finally, 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,[29] you can use the Data Catalog search capabilities to easily search across databases and tables within your data lake. This feature allows 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. Thus, if a consumer can be identified in more than one way (i.e., a user ID, group ID, etc.) using these search capabilities will help to 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 set-up and allows 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, you simply sign in to 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 AWS CloudTrail page.[30] 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.[31]

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 Amazon S3 buckets, table updates in Amazon 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 else required to run 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.[32] This Lambda function may contain code to gather and delete all metadata associate with a given identifier provided by the consumer. You also may 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 could be used to obtain customer information 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.[33]

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.

AWS Config

AWS Config[34] is a continuous monitoring and assessment service that records changes to the configuration of your AWS resources. You can view the current and historic configurations of a resource and use this information to troubleshoot outages, conduct security attack analysis, and much more. You can view the configuration at any time and use that information to re-configure your resources and return them to a steady state in an outage situation.

Using Config Rules, you can run continuous assessment checks on your resources to verify that they comply with your own security policies, industry best practices, and compliance requirements such as PCI and HIPAA. For example, AWS Config provides managed Config Rules to ensure that encryption is enabled for all EBS volumes in your account. You also can write a custom Config Rule to essentially “codify” your own security policies. AWS Config alerts you in real time when a resource is misconfigured, or when a resource violates a particular security policy. Specifically, in regard to the CCPA, you can create custom Config Rules to monitor your resources. For example, you can use the "required-tags" identifier to check whether your resources have the tags that you require for collection purposes.

Amazon SES

Amazon Simple Email Service (Amazon SES)[35] 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[36] 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 analytics tools.

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.

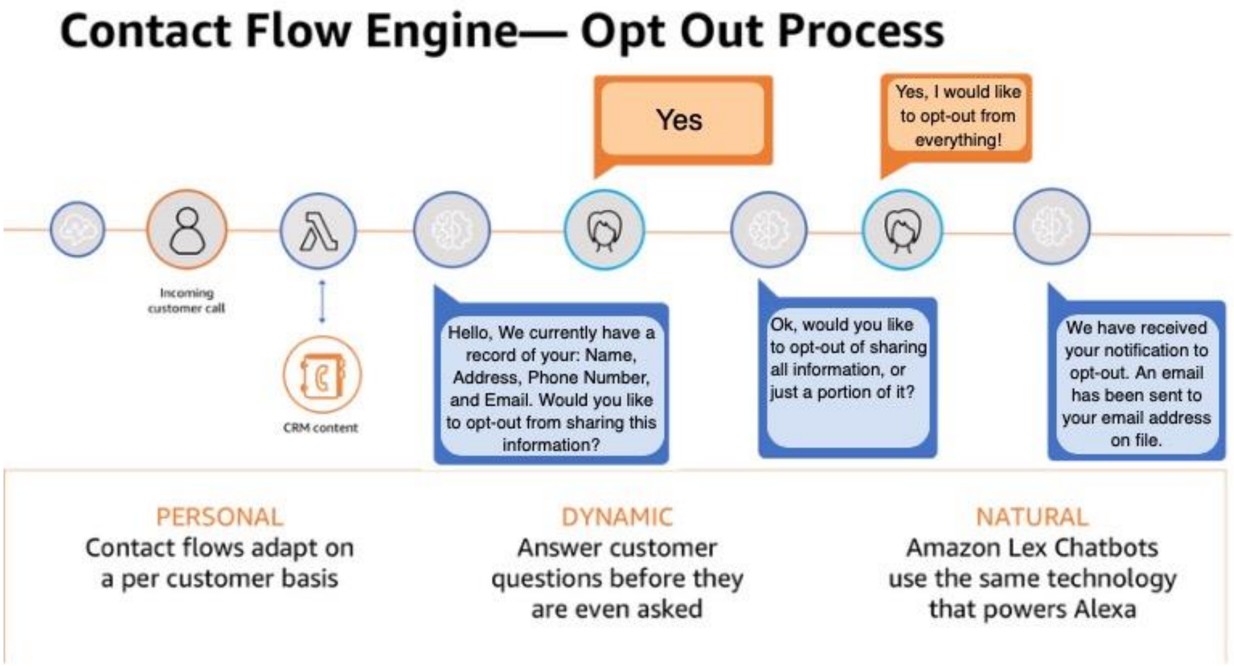
The contact flows in Amazon Connect can be used to create dynamic interactive voice response (IVR) solutions. With Amazon Connect, you can gather appropriate personal information to customize your customer’s experience when they interact with your IVR. The personal information used for this customization can include social security numbers, credit card information, and addresses. For compliance reasons, sensitive personal information must be encrypted while in motion and when stored. Always encrypt personal information.

Amazon Connect, when paired with Amazon Lex[37] 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 as 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 set it up 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, and automatically encrypt that data using your encryption keys. This feature allows you to 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 protects 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[38].

You can use AI Services from AWS[39] 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.


Contact Flow Engine— Opt Out Process
Incoming customer call
CRM content
PERSONAL
Contact flows adapt on a per customer basis
DYNAMIC
Answer customer questions before they are even asked
NATURAL
Amazon Lex Chatbots use the same technology that powers Alexa

Hello, We currently have a record of your: Name, Address, Phone Number, and Email. Would you like to opt-out from sharing this information?

Yes

Ok, would you like to opt-out of sharing all information, or just a portion of it?

Yes, I would like to opt-out from everything!

We have received your notification to opt-out. An email has been sent to your email address on file.

In the process shown in the figure above, a customer dials a company's toll-free number, and their phone number is detected by Amazon Connect. which 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).

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 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 agent-less 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[40] to learn about options for building anything from small scripts that delete data to a full orchestration framework that uses AWS Code services.[41]

Contributors

Contributors to this document include:

- Julia Soscia, 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

Further Reading

For additional information, see:

- AWS Best Practices for DDoS Resiliency[42]
- AWS Security Checklist[43]
- Securing Data at Rest with Encryption[44]
- Cloud Adoption Framework - Security Perspective[45]
- Introduction to AWS Security Processes[46]
- AWS Security Best Practices[47]
- Encrypting Data at Rest[48]
- AWS Risk and Compliance[49]
- Using AWS in the Context of Common Privacy and Data Protection Considerations[50]
- Security at Scale: Logging in AWS[51]
- Security at Scale: Governance in AWS[52]
- Secure Content Delivery with Amazon CloudFront[53]

Document Revisions

Date	Description
July 2019	First publication

[1] https://aws.amazon.com/compliance/programs/

[2] https://d1.awsstatic.com/whitepapers/Security/AWS_Security_Best_Practices.pdf

[3] https://aws.amazon.com/privacy/

[4] https://aws.amazon.com/compliance/shared-responsibility-model/

[5] https://aws.amazon.com/compliance/iso-27001-faqs/

[6] https://aws.amazon.com/compliance/iso-27017-faqs/

[7] https://aws.amazon.com/compliance/iso-27018-faqs/

[8] https://aws.amazon.com/compliance/iso-9001-faqs/

[9] https://aws.amazon.com/compliance/pci-dss-level-1-faqs/

[10] https://aws.amazon.com/compliance/soc-faqs/

[11] https://aws.amazon.com/compliance/programs/

[12] https://d1.awsstatic.com/whitepapers/Security/AWS_Security_Whitepaper.pdf

[13] https://aws.amazon.com/artifact/

[14] https://aws.amazon.com/s3/

[15] https://docs.aws.amazon.com/AmazonS3/latest/API/RESTObjectPUTtagging.html

[16] https://docs.aws.amazon.com/AmazonS3/latest/API/RESTObjectGETtagging.html

[17] https://docs.aws.amazon.com/AmazonS3/latest/API/RESTObjectDELETEtagging.html

[18] https://aws.amazon.com/lambda/

[19] https://aws.amazon.com/api-gateway/

[20] https://aws.amazon.com/dynamodb/

[21] https://aws.amazon.com/kms/

[22] https://aws.amazon.com/rds/

[23] https://aws.amazon.com/redshift/

[24] https://aws.amazon.com/emr/

[25] https://aws.amazon.com/glue/

[26] https://aws.amazon.com/athena/

[27] https://aws.amazon.com/quicksight/

[28] https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-hive-metastore-glue.html

[29] https://aws.amazon.com/lake-formation/

[30] https://aws.amazon.com/cloudtrail/

[31] https://d0.awsstatic.com/whitepapers/compliance/AWS_Security_at_Scale_Logging_in_AWS_Whitepaper.pdf

[32] https://docs.aws.amazon.com/AmazonCloudWatch/latest/events/RunLambdaSchedule.html

[33] https://media.amazonwebservices.com/architecturecenter/AWS_ac_ra_web_01.pdf

[34] https://aws.amazon.com/config/

[35] https://aws.amazon.com/ses/

[36] https://aws.amazon.com/connect/

[37] https://aws.amazon.com/lex/

[38] https://docs.aws.amazon.com/encryption-sdk/latest/developer-guide/how-it-works.html#envelope-encryption

[39] https://aws.amazon.com/machine-learning/ai-services/

[40] https://aws.amazon.com/tools/

[41] https://aws.amazon.com/products/developer-tools/

[42] https://d0.awsstatic.com/whitepapers/Security/DDoS_White_Paper.pdf

[43] https://d0.awsstatic.com/whitepapers/Security/AWS_Security_Checklist.pdf

[44] https://d0.awsstatic.com/whitepapers/AWS_Securing_Data_at_Rest_with_Encryption.pdf

[45] https://d0.awsstatic.com/whitepapers/AWS_CAF_Security_Perspective.pdf

[46] https://d0.awsstatic.com/whitepapers/Security/Intro_Security_Practices.pdf

[47] https://d0.awsstatic.com/whitepapers/Security/AWS_Security_Best_Practices.pdf

[48] https://d0.awsstatic.com/whitepapers/AWS_Securing_Data_at_Rest_with_Encryption.pdf

[49] https://d0.awsstatic.com/whitepapers/compliance/AWS_Risk_and_Compliance_Whitepaper.pdf

[50] https://d0.awsstatic.com/whitepapers/compliance/Using_AWS_in_the_context_of_Singapore_Privacy_Considerations.pdf

[51] http://d0.awsstatic.com/whitepapers/compliance/AWS_Security_at_Scale_Logging_in_AWS_Whitepaper.pdf

[52] http://d0.awsstatic.com/whitepapers/compliance/AWS_Security_at_Scale_Governance_in_AWS_Whitepaper.pdf

[53] https://d0.awsstatic.com/whitepapers/Security/Secure_content_delivery_with_CloudFront_whitepaper.pdf