Data Classification

Secure Cloud Adoption

March 2020

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

Data Classification Overview 1
Data Classification Value 1
Data Classification Process 2
Existing Data Classification Models 3
U.S. National Security Classification Scheme 4
U.S. Information Categorization Scheme 5
United Kingdom (UK) Data Classification Scheme 5
Customer Considerations for Implementing Data Classification Schemes 6
Data Classification and Privacy Considerations 7
Newer Considerations in Data Classification 7
AWS Recommendations 8
Enterprise Approaches 10
Leveraging AWS Cloud to Support Data Classification 12
Document Revisions 14

Abstract

This paper provides insight into data classification categories for public and private organizations to consider when moving data to the cloud. It outlines a process through which customers can build data classification program, shares examples of data and the corresponding category it may fall into, and outlines practices and models currently implemented by global first movers and early adopters along with data classification and privacy considerations. It also examines how implementation of data classification program can simplify cloud adoption and management, and recommends that customers leverage internationally recognized standards and frameworks when developing their own data classification rules.

Data Classification Overview

Data classification is a foundational step in cybersecurity risk management. It involves identifying the types of data that are being processed and stored in an information system owned or operated by an organization. It also involves making a determination on the sensitivity of the data and the likely impact should the data face compromise, loss, or misuse.

To ensure effective risk management, organizations should aim to classify data by working backwards from the contextual use of the data and creating a categorization scheme that takes into account whether a given use-case results in significant impact to an organization's operations (e.g. if data is confidential, needs to have integrity, and/or be available).

> As used in this document, the term "classification" implies a holistic approach inclusive of taxonomy, schemes, and categorization of data for confidentiality, integrity, and availability.

Data Classification Value

Data classification has been used for decades to help organizations make determinations for safeguarding sensitive or critical data with appropriate levels of protection. Regardless of whether data is processed or stored in on premise systems or the cloud, data classification is a starting point for determining the appropriate level of controls for the confidentiality, integrity, and availability of data based on risk to the organization. For instance, data that is considered "confidential" should be treated with a higher standard of care than "public" data consumed by the general public. Data classification allows organizations to evaluate data based on sensitivity and business impact, which then helps the organization assess risks associated with different types of data. Standards organizations, such as the International Standards Organization (ISO) and the National Institute of Standards and Technology (NIST), recommend data classification schemes so information can be effectively managed and secured according to its relative risk and criticality, advising against practices that treat all data equally. Each data classification level should be associated with a recommended baseline set of security controls that provide protection against vulnerabilities, threats, and risks commensurate with the designated protection level.

It is important to note the risks with over classifying data. Sometimes organizations err by broadly classifying large disparate sets of data at the same sensitivity level. This over-classification can incur unwarranted expenses by putting into place costly controls that can additionally impact business operations. This approach can also divert attention to less critical datasets and limit business use of the data through unnecessary compliance requirements due to over classification.

Data Classification Process

Customers often seek tangible recommendations when it comes to establishing data classification policies. These steps help not only in the development phase but can be used as measures when reassessing if datasets are in the appropriate tier with corresponding protections.

The paragraphs below provide a step-by-step approach, based on internationally-recognized guidance that customers can consider when developing data classification policies[1][2]:

1. *Establishing a data catalog:* Conducting an inventory of the various data types that exist in the organization, how is it used, and if any of it is governed by a compliance regulation or policy. Once the inventory is complete, group the data types into one of the data classification levels the organization has adopted.
2. *Assessing business critical functions and conduct an impact assessment*: An important aspect in determining the appropriate level of security for data sets is to understand the criticality of that data to the business. Following an assessment of business critical functions, customers can conduct an impact assessment for each data type.
3. *Labeling information:* Undergo a quality assurance assessment to ensure that assets and data sets are appropriately labeled in their respective classification buckets. Additionally, it may be necessary to create secondary labels for data sub-types to differentiate particular sets of data within a tier due to privacy or other compliance concerns. Using services like Amazon SageMaker and AWS Glue provide insight and can support in data labeling activities.

[1] ISO 27001/27002 is a widely-adopted global security standard that sets out requirements and best practices for a systematic approach to managing company and customer information that's based on periodic risk assessments appropriate to ever-changing threat scenarios

[2] https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-60v1r1.pdf

4. *Handling of assets:* When data sets are assigned a classification tier, data is handled according to the handling guidelines appropriate for that level, which include specific security controls. These handling procedures should be formalized but also adjust as technology changes. (Refer to "Customer Considerations for Implementing Data Classification Schemes" below for additional information on data handling.
5. *Continuous monitoring:* Continue to monitor the security, usage and access patterns of systems and data. This can be done through automated (preferred) or manual processes to identify external threats, maintain normal system operations, install updates, and track changes to the environment.

Existing Data Classification Models

The United States (U.S.) and the United Kingdom (UK) have established data classification schemes for public sector data. Both governments use a three-tiered classification scheme with the majority of public sector data classified in the two lowest tiers. It's important to note that for some governments, more extensive data classification may be useful. For example, the city of Washington, D.C. in the United States, has established a data classification program using a five-tiered classification scheme that was widely applauded by open data advocates, and may be a good model for other local governments. Data classification schemes have a short list of attributes and associated measures or criteria that help organizations determine the appropriate categorization level.

The city of Washington, D.C. implemented a new data policy in 2017 focused on being more transparent, while still protecting sensitive data. While Washington D.C. implemented a five tier model, these tiers can align with other widely-adopted three-tier classification schemes used in cloud accreditation regimes.[3]

Level 0 — Open Data. Data readily available to the public on open government websites and datasets.

Level 1 — Public Data, Not Proactively Released. Data not protected from public disclosure or subject to withholding under any law, regulation, or contract. Publication of the data on the public Internet would have the potential to jeopardize the safety, privacy, or security of anyone identified in the information.

3 https://octo.dc.gov/page/district-columbia-data-policy

Level 2 — For District Government Use. Data that is not highly sensitive and may be distributed within the government without restriction by law, regulation, or contract. It is primarily daily government business operations data.

Level 3 — Confidential. Data protected from disclosure by law, regulation, or contract and that is either highly sensitive or is lawfully, regulatory, or contractually restricted from disclosure to other public bodies. This includes privacy-related data (e.g., personally identifiable information (PII), protected health information (PHI), payment card industry data security standard (PCI DSS), federal tax information (FTI), etc.)

Level 4 — Restricted Confidential. Data that unauthorized disclosure could potentially cause major damage or injury, including death to those identified in the information, or otherwise significantly impair the ability of the agency to perform its statutory functions.

U.S. National Security Classification Scheme

The U.S. government uses a three-tier classification scheme for national security information as described in Executive Order 135261. This scheme is focused on handling instructions based on potential impact to national security if it is disclosed (i.e. confidentiality).

1. Confidential — Information where unauthorized disclosure reasonably could be expected to cause damage to national security.

2. Secret — Information where unauthorized disclosure reasonably could be expected to cause serious damage to national security.

3. Top Secret — Information where unauthorized disclosure reasonably could be expected to cause exceptionally grave damage to national security.

Within these classification tiers there are also secondary labels that can be applied that give origination information and can modify the handling instructions. The U.S. also uses the term “unclassified data” to refer to any data that is not classified under the three classification levels. Even with unclassified data, there is the potential use of secondary labels for sensitive information, such as “For Official Use Only” (FOUO) and “Controlled Unclassified Information” (CUI) that restrict disclosure to the public or unauthorized personnel.

U.S. Information Categorization Scheme

Due to the targeted focus of the U.S. classification system and to address additional risks to information beyond confidentiality, NIST developed a three-tiered categorization scheme based on the potential impact to the confidentiality, integrity, and availability of information and information systems applicable to an organization's mission. Most of the data processed and stored by public sector organizations can be categorized into the following:

NIST developed a three-tiered categorization scheme based on the potential impact to the confidentiality, integrity, and availability of information and information systems.

• Low — limited adverse effect on organization operations, organization assets, or individuals.

• Moderate — serious adverse effect on organization operations, organization assets, or individuals.

• High — severe or catastrophic adverse effect on organization operations, organization assets, or individuals.

According to Fiscal Year 2015 data[4], U.S. federal departments and agencies categorized 88 percent of their systems into the low and moderate categories. AWS has regions and services that are accredited to support all types of data categories and classifications.

United Kingdom (UK) Data Classification Scheme

In 2014, the UK simplified its data classification scheme by reducing the levels from six to three. They are:

1. Official — Routine business operations and services, some of which could have damaging consequences if lost, stolen, or published in the media, but none of which is subject to a heightened threat profile.

[4] https://www.gao.gov/assets/710/700588.pdf

2. Secret — Very sensitive information that justifies heightened protective measures to defend against determined and highly capable threat actors (e.g., compromise could significantly damage military capabilities, international relations, or the investigation of serious organized crime).
3. Top secret — Most sensitive information requiring the highest levels of protection from the most serious threats (e.g., compromise could cause widespread loss of life or could threaten the security or economic well-being of the country or friendly nations).

According to a cabinet office core briefing, the UK government categorized approximately 90 percent of its data as "Official"[5], which serves as the basic level of data classification, followed by 'secret' and 'top secret'. The UK uses a flexible, de-centralized accreditation approach where individual agencies determine the cloud services suitable for "Official" data based on a cloud service provider's (CSP's) security assurance against 14 cloud security principles[6]. Most UK government agencies have determined that it is appropriate to use reputable, hyper-scale CSPs when running workloads with "Official" data.

Customer Considerations for Implementing Data Classification Schemes

In addition to implementing a data classification scheme, it is equally important to determine data handling roles. ISO, NIST, and other standards place the responsibility of data classification on data owners, as they are the best positioned to determine the value, use, sensitivity, and criticality of their own data. Risk management obligations vary depending on the role of the parties that handle the data. In other words, data owners (i.e. controllers who generate and control content, such as agencies and ministries) and non-data owners (i.e. processors that handle data in order to provision services) should be subject to requirements appropriate for the roles they play. In the context of public sector data classification, agencies or ministries work as the data

[5] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/251481/Government-Security-Classifications-Supplier-Briefing-Oct-2013.pdf

[6] https://www.ncsc.gov.uk/collection/cloud-security?curPage=/collection/cloud-security/implementing-the-cloud-security-principles

owner and are responsible for classifying their data and determining the security accreditation that they expect their CSP to meet.

It is important to note that organizations applying a blanket high classification level to all data (despite its true risk posture) do not reflect a risk-based, outcome-focused approach to security. Protecting data classified at higher levels requires a higher standard of care, which translates into the customer spending increased resources on securing, monitoring, measuring, remediating, and reporting risks. It is impractical to commit the significant resources required to securely manage higher impact data for data that does not meet the requisite thresholds. Also, the additional controls placed on data at the lower classification levels can negatively affect the availability, completeness or timeliness of that data to the general workforce, customers, and/or constituents. Where risks can be managed so that data is handled at a lower classification level, organizations will experience the most flexibility around how they use that data.

Data Classification and Privacy Considerations

Data classification is particularly important as new global privacy laws and regulations provide consumers with rights to access, deletion, and other controls over personal data. For instance, under the European Union's General Data Protection Regulation, organizations are required to respond to certain consumer requests within a month of receipt. In order to respond appropriately, organizations must generally verify a requester's identity, locate the requestor's personal data, ensure the data returned only contains the requestor's personal data, and possibly refuse a request if it's inconsistent with applicable law. Organizations that adopt strong data classification policies are better positioned to provide timely responses to these requests. A data classification framework along with proper tagging and labeling will help protect this personal data. Secondary labels can be used within a classification tier to assist tagging and discovery of relevant privacy data. This allows an organization to quickly address issues as they arise. Such additional mechanisms also aid in traceability and access monitoring of sensitive data sets.

Newer Considerations in Data Classification

Whether the journey to cloud is nascent or established, it's critical to establish data classification rules. Similar to reviewing existing security practices and establishing better policies based on newer threats, considerations in how to protect data are highlighted in this section as an example of what customers should consider when

revisiting existing data classification policies. Most recently, conversations in industry consortiums have raised the following points:

1. **Data is scattered everywhere:** The ubiquitous use of modern technology and reliance on information in enterprises across all sectors means massive volumes of data are stored, processed, and in transit across numerous systems, devices, and end users. This can pose significant challenges for enterprises that are responsible for managing and securing large volumes of data.
2. **Intra- and inter-organizational dependencies:** The ever increasing need to collaborate and share information within an organization and across organizations within the same sector or with similar mission needs (e.g., hospital and health care networks).
3. **End user knowledge:** Models that rely on end users to identity and classify data, such as those for machine learning processes, can be error prone and often incomplete. End-users may lack the skills or awareness of risks to categorize and manage data effectively.
4. **Data classifiers and tagging:** There is usually a lack of common definitions and understanding of classifiers, along with a lack of standards across industries or persistence of labeling
5. **Context:** Context matters. The actual sensitivity and criticality of information depends greatly on other factors, such as how it is used and with whom, than what the information is necessarily about.

While these challenges may not seem new, they are factors worth considering as organizations develop and implement data classification.

AWS Recommendations

In most cases, AWS recommends starting with a three-tiered data classification approach (Table 1), which has shown to sufficiently meet both public and commercial customer needs and requirements. As an example, the table below includes three tiers and a naming convention for each tier. For organizations that have more complex data environments or varied data types, secondary labeling is helpful without adding complexity with more tiers. We recommend using the minimal number of tiers that makes sense for the organization.

Table 1: Three-tiered data classification approach

Data Classification	System Security Categorization	Cloud Deployment Model Options
Unclassified	Low to High	Accredited public cloud
Official	Moderate to High	Accredited public cloud
Secret and above	Moderate to High	Accredited private/hybrid/community cloud/ public cloud

Data Residency Consideration: AWS encourages customers to assess their data classification approach and hone in on which data needs to stay within their country or region, and why. By doing so, customers may find that their data, potentially even sensitive and critical data, may be stored and/or replicated elsewhere if there is no particular legal or policy geographical requirement. This can further reduce risk of loss in the event of a disaster and provide access to technologies and capabilities that may not be available in their area. Learn more in the AWS Data Residency whitepaper.

NIST's data classification scheme has been widely recognized in sector-specific, national and international certifications. In fact, governments such as the Philippines and Indonesia are evaluating and adopting data classification schemes that apply similar principles as the US and UK models. However, organizations are best positioned to develop their own classification schemes based on organizational and risk management needs. Organizations seeking to move away from heavy, more burdensome tiered schemes can execute risk impact assessments and then move forward with fewer tiered schemes that are easier to manage and classify, such as the three-tiered model.

Organizations should select the appropriate cloud deployment model according to their specific needs, the type of data they handle, and assessed risk (refer to table below). Depending on the classification of the data, they will need to apply the relevant security controls (e.g., encryption) within their cloud environment.

When assessing risk and determining security controls, it is important to understand how commercial cloud services differ from on-premises systems, the differences in implementation of controls (i.e., shared responsibility model), and that there may be

alternate controls to consider as compared to traditional IT implementation. When organizations have fully evaluated the commercial cloud with the numerous security benefits available (e.g., improved availability and resiliency, improved visibility and automation, and continually audited infrastructure), they may find that the vast majority of their workloads can be deployed in the cloud with due regard to a data classification scheme, similar to what the US and UK governments have done. Globally, we are seeing public sector organizations increasingly leverage the native security benefits of commercial cloud and meeting their security and compliance requirements through appropriate data classification and implementation of security controls.

When organizations have fully evaluated the commercial cloud with its numerous security benefits, they may find that the vast majority of their workloads can be deployed in the cloud with due regard to a data classification scheme, similar to what the US and UK governments have done.

Enterprise Approaches

This section identifies industry-specific examples for data classification, which may include sector-specific requirements. As mentioned earlier, different data types (e.g., government, financial, and healthcare data) may require additional considerations for tiers and secondary labels to address different handling procedures. Regardless of data belonging to public or commercial entities, customers must conduct the due diligence of adhering to local compliance and regulatory requirements.

The following chart contains examples of data classification schemes in practice today, descriptions of what can be included in that category based on tier, and examples of workload types for a particular tier.

Example 1

Data Classification	Examples of Workloads
Tier 3 – Government confidential and above-sensitive data	National security and defense information Government intelligence information Law enforcement information Government program monitoring or oversight investigations information

Data Classification	Examples of Workloads
Tier 2 – Restricted	Personally identifying information about individuals Human Resources Management Personal profile information Aggregated financial or market data
Tier 1 – Public data	Marketing or promotional information Information related to other general government administrative or program activities Intra-agency workplace policy development and management

Example 2

Data Classification	Examples
Tier 3 – Highly Strategic	Highly sensitive trade secret and material confidential business information (e.g., certain pricing, merger/acquisition information, marketing plan, proprietary processes, marketing plans, new product designs, inventions prior to a patent application or held as trade secret) the public disclosure of which could be expected to cause severe or catastrophic legal, financial or reputational damage.
Tier 2 – Restricted	Most material and non-material business data (e.g., email, sales and marketing account data, executed contracts, receipts) Information required by law to be protected from unauthorized disclosure Employee HR records (including employee disciplinary reports)
Tier 1 – Protected data	CRM systems Vendor bank account numbers and payment instructions Information that is available only to a specific group of the company's employees for the purpose of conducting business Information for only internal use

Leveraging AWS Cloud to Support Data Classification

Cloud computing can offer customers the ability to secure their workloads; whether in highly regulated industries, public sector, or small-medium sized businesses, to meet data classification policies and requirements. Cloud service providers (CSPs), such as AWS, provide a standardized, utility-based service that is self-provisioned by customers. CSPs do not have visibility into the type of data customers run in the cloud, which means CSPs do not distinguish, for example, personal data from other customer data when providing cloud services. It is the customer's responsibility to classify their data and implement appropriate controls within their cloud environment (e.g., encryption). However, the security controls CSPs implement within their infrastructure and their service offerings can be used by customers to meet the most sensitive data requirements.

AWS services offer the same high level of security to all customers, regardless of the type of content being stored. AWS adopts a high security bar across all services. These services are then queued for certification against international security and compliance "gold" standards, which translates to customers benefiting from elevated levels of protection for customer data processed and stored in the cloud. The risk events and threat vectors of greatest concern are largely accounted for through foundational cyber hygiene disciplines (e.g., patching and configuring systems), which CSPs can demonstrate through widely adopted, internationally-recognized security certifications such as ISO 27001[7], Payment Card Industry Data Security Standard (PCI DSS)[8], and Service Organization Controls (SOC)[9]. In evaluating CSPs, customers should leverage these existing CSP certifications so that the customer can appropriately determine whether a CSP (and services within the CSP's offerings) can support their data classification requirements. We encourage organizations to implement a policy identifying which existing national, international, or sector-specific cloud certifications

[7] ISO 27001/27002 is a widely-adopted global security standard that sets out requirements and best practices for a systematic approach to managing company and customer information that's based on periodic risk assessments appropriate to ever-changing threat scenarios

[8] The Payment Card Industry Data Security Standard (also known as PCI DSS) is a proprietary information security standard administered by the PCI Security Standards Council (https://www.pcisecuritystandards.org/), which was founded by American Express, Discover Financial Services, JCB International, MasterCard Worldwide and Visa Inc. PCI DSS applies to all entities that store, process or transmit card

[9] Service Organization Controls reports (SOC 1, 2, 3) are intended to meet a broad range of financial auditing requirements for U.S. and international auditing bodies. The audit for this report is conducted in accordance with the International Standards for Assurance Engagements No. 3402 (ISAE 3402) and the American Institute of Certified Public Accountants (AICPA): AT 801 (formerly SSAE 16).

and attestations are acceptable for each level in the data classification scheme to streamline accreditation and accelerate migrating workloads to the cloud.

AWS offers several services and features that can facilitate an organization's implementation of a data classification scheme. For example, Amazon Macie can help customers inventory and classify sensitive and business-critical data stored in AWS. Amazon Macie uses machine learning to automate the process of discovering, classifying, labeling, and applying protection rules to data. This helps customers better understand where sensitive information is stored and how it's being accessed, including user authentications and access patterns.

Other AWS services and features that can support data classification include, but are not limited to:

- AWS Identity and Access Management (IAM) for managing user credentials, setting permissions, and authorizing access.
- AWS Organizations helps you centrally govern your environment with automated account creation, accounts grouping to reflect your business needs, and policies to enforce governance. Policies can include required actions such as tagging of resources
- AWS Glue to store data and discover associated metadata like table definition and schema, in the AWS Glue Data Catalog. Once cataloged, your data is immediately searchable and available for ETL.
- Amazon Neptune, fully managed graph database, can give customers insights into the relationships between different data sets. This can include identification and traceability of sensitive data through metadata analysis.
- AWS KMS or AWS CloudHSM for encryption key Management with AWS-generated keys or bring your own key (BYOK) with FIPS 140-2 validation.
- AWS CloudTrail for extensive logging to track who, what, and when data was created, accessed, copied/ moved, modified, and deleted.
- AWS Systems Manager to view and manage service operations like patching along with AWS Inspector to conduct vulnerability scans.
- AWS GuardDuty for intelligent threat detection supporting continuous monitoring requirements.
- AWS Config to manage configuration changes and implement governance rules.

- AWS Web Application Firewall (WAF) and AWS Shield to protect web applications from common attack vectors (e.g., SQL Injection, Cross-Site Scripting, and DDoS).

To review the entire list of AWS security services, see Security, Identity, and Compliance on AWS.

Document Revisions

Date	Description
March 2020	Updated to reflect latest services and technologies.
June 2018	First publication