Data Sheet

CipherTrust Data Discovery and Classification FAQs

cpl.thalesgroup.com

THALES
Building a future we can all trust

Contents

THALES
Building a future we can all trust

Business and Market Drivers

What is CipherTrust Data Discovery and Classification (DDC)?

CipherTrust Data Discovery and Classification provides complete visibility into the location of sensitive data across your enterprise, so you can uncover and close compliance gaps. DDC scans structured as well as unstructured data stores for named entities in different formats and global languages to help you find any type of sensitive data, in any language, anywhere across your enterprise. Once you have identified security blind spots you can quickly remediate using one of the CipherTrust Platform's market-leading encryption solutions.

What customer problems/use-cases does it address?

Compliance with security and privacy regulations: Organizations need to protect Personally Identifiable Information (PII) against data leaks and improper use to comply with the requirements of privacy regulations like General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), Health Insurance Portability and Accountability Act (HIPAA), and Brazilian General Data Protection Law (LGPD). CipherTrust DDC provides complete visibility into the location of sensitive data, enhancing an organization's ability to adopt appropriate data security controls and measure to protect sensitive personal data from loss and unauthorized access.

Increase data visibility: Organizations need greater data visibility to support better decision making for risk analysis and remediation, as well as reporting and compliance. According to the 2024 Thales Data Threat Report, 70% of enterprises are able to classify only 50% or less of their data. CipherTrust DDC automatically scans your data stores across on-premises, hybrid, and multicloud environments to help you protect and manage your sensitive data.

Reduce exposure risk during cloud migration: Major change programs like digital transformation involve moving large amounts of sensitive data from one environment to another. Uncontrolled dispersal of data across cloud platforms increases the potential of a data breach event, as well as infringement of privacy regulations. As IT environments become more complex, it becomes more difficult to discover sensitive data and have oversight or manage access across data sources. CipherTrust DDC provides visibility into exactly what information you have stored so you can plan an effective strategy for transformation to safeguard data at each stage of the process.

Secrets discovery: Modern development trends like containerization, DevOps and automation have contributed to a massive increase in the use of secrets (credentials, certificates, keys) for authentication. Secrets can be vulnerable to cyberattacks when not securely managed. The CipherTrust Data Security Platform provides a simplified workflow to address this risk. DDC automatically discovers more than 30 different types of secrets, including AES Keys, Auth Secrets, and SSH Keys. Once exposed secrets have been discovered, security teams can take actions to remediate the risk and improve security posture using CipherTrust Secrets Management.

What are the industries that use and need CipherTrust Data Discovery and Classification?

Highly regulated industries such as financial, insurance, healthcare, manufacturing, retail, and government are the target market for this solution. In fact any industry that needs to comply with privacy regulations, especially if they are undergoing digital transformation, sharing sensitive data with an ecosystem of business partners or suppliers, or tracking sensitive data at endpoints - all are target customers for CipherTrust Data Discovery and Classification.

How does the solution help protect data and enforce compliance?

CipherTrust Data Discovery and Classification efficiently locates sensitive data across an enterprise using a streamlined workflow that automates discovery, classification and protection, eliminates security blind spots, and provides a clear view of the sensitive data and its risks. It comes with a comprehensive set of built-in templates for rapid discovery of regulated data. As a result, organizations can more easily uncover and close their data protection gaps, prioritize their remediation efforts, and proactively respond to a growing number of data privacy and data security regulations.

Fundamental Concepts

What is data discovery?

Data discovery is the process of finding sensitive or regulated data, both structured and unstructured, stored across different data stores, for example, the cloud, file servers, databases, big data, devices, backups, snapshots, etc.. Data discovery is helpful in finding sensitive data that customers need to protect in order to comply with various regulations. Given the volume and variability of data and data stores, this can be challenging for any organization.

What is data classification?

Data classification is the process of categorizing data based on pre-defined criteria, e.g., built-in templates for data privacy regulations or custom profiles created by an organization. Data classification helps determine the appropriate levels of security necessary for sensitive data. CipherTrust Data Discovery and Classification provides four levels of classification by default. These classification levels are built-in with the following descriptions. If an organization wants to add more levels, or change the definition of any level, they are free to edit the classification levels.

- Restricted: This is highly sensitive data, e.g., customer personal data, trade secrets, etc., requiring the best possible data security. Disclosure of such data can lead to severe financial and legal consequences for an organization. Businesses must prioritize remediation efforts related to this type of data.
- Private: Disclosure of this data can cause serious damage to an organization. While it is less sensitive than restricted, it requires a high level of protection.
- Internal: This is data with low sensitivity, e.g., draft of a planning document. Exposure of such data may not affect an organization very much. An important consideration is that the data generally is not intended for public disclosure.
- Public: This is the least sensitive data with no specific need for data security, e.g., press releases and marketing collateral which are published in the public domain on web sites. Such data can be freely shared with external entities.

What is data remediation?

Data remediation is the process of mitigating the risks of data exposure so that sensitive data remains protected from unauthorized access and use. This can be achieved, using any one or a combination of data protection techniques such as data encryption, tokenization, identity and access controls, and other methods.

How does DDC compare with Data Loss Prevention (DLP) solutions?

The DLP solutions focus on preventing sensitive data from leaving the organization's perimeter. CipherTrust Data Discovery and Classification focuses on data privacy and protection - identifying sensitive data, and getting a clear understanding of data and its risk. This enables organizations to take appropriate steps to protect their data and comply with data privacy and data security regulations.

Solution Implementation Details

What are the deployment options?

The solution is deployed on-premises by installing an agent on the host, or remotely via a proxy agent.

What are the pros and cons for agent-based and agentless deployments?

Below are the recommendations for agent-based and agentless deployments:

Both approaches share the following pros:

Type of Agent	Value	Recommendation
Agent-based	• Information does not need to be transmitted over the network to be scanned. • Faster scan • No need for credentials for scanning	Scanning data stores that allow an Agent to be installed locally. E.g. local storage and local memory on server or workstation with installed Agents
Agentless	• Faster deployment, as Agents do not have to be installed directly on Target hosts • Can scan multiple targets • It can scan any type of targets • It doesn't consume resources on the target host	Scanning data stores that can only be accessed remotely. E.g. database systems, email servers, cloud storages and network storage locations

Where do the discovery scans run?

The discovery scans run locally, next to the data location.

Is an agent required for each data store?

No. If you have multiple data stores on the same host, you can use one agent to scan all the targets. The proxy agents can also scan multiple data stores.

How is an agent selected for scanning?

The agent selection for scanning a data store is done automatically by CipherTrust Data Discovery and Classification. The status of this process is shown in the summary page for the data stores. The agent selection is done during the data store creation process and the customer can select a specific agent using the label capability and the number of agents for scanning the target location. Depending on the data store type this capability could be available or not, for example local storages don't have any of these capabilities as they can only be scanned using the local agent.

Is it possible to scan a specific path or table within a data store?

Yes, this can be configured during the scan creation. Once the data store to scan has been selected, the customer can select a specific target to scan, for example a table within a database or a specific path within a local storage.

What factors should be considered in deciding how many agents are necessary?

CipherTrust Data Discovery and Classification can discover one or more data stores (file servers, DBs, network storages, big data, etc.) in a single scan. While it is hard to generalize, here are some considerations:

- The amount of data to be scanned in each data store. For example, a client can use multiple agents to scan a Hadoop data store, each agent scanning a different path.
- The frequency and total number of scans to be executed.

Can customers define their own types of data?

Yes. The customers can define their own type of data based on patterns and regular expressions.

What countries does the solution cover?

The current version of the solution covers data types for countries in various regions that can be used in scans.

Region	Countries	
Africa	• Gambia • South Africa	
Asia	• Hong Kong • Japan • Malaysia • People's Republic of China • Singapore	• South Korea • Sri Lanka • Taiwan • Thailand
Europe	• Austria • Belgium • Bulgaria • Croatia • Cyprus • Czech Republic • Denmark • Finland • France • Germany • Greece • Hungary • Iceland • Ireland • Italy • Latvia	• Luxembourg • Macedonia • Malta • Netherlands • Norway • Poland • Portugal • Romania • Serbia • Slovakia • Slovenia • Spain • Sweden • Switzerland • Turkey • United Kingdom • Yugoslavia (former)
Middle East	• Iran • Israel • Saudi Arabia • United Arab Emirates	
North America	• Canada • Mexico • United States of America	
Oceania	• Australia • New Zealand	
South America	• Brazil • Chile	

Does DDC provide support for public cloud platforms?

Yes.

How can an organization see the results of a scan?

CipherTrust Data Discovery and Classification enables organizations to get a clear understanding of their sensitive data, locations, risks, and the protection status from a centralized console. The users will be able to access detailed reports for audits and risks mitigation, all from a centralized console.

What do the risk scores signify and how are they used?

The risk scores allow organizations to identify the sensitivity of data objects like files and databases by aggregating various parameters such as protection level, number of elements found, location, amount of sensitive data, etc. With risk scores, businesses can identify the sources that are at most risk and take actions to protect sensitive data. In the future, risk scores will provide additional information such as average risk score so organizations can compare the risk posed by each data asset.

Which compliance templates are included in the solution?

Some examples: APPI, CCPA, GDPR, HIPAA, NDB, PCI DSS, LGPD, UK-GDPR and SHIELD.

What languages are supported by the solution?

CipherTrust DDC can scan any Unicode-based characters, meaning almost all the languages. From a GUI point of view, CipherTrust DDC supports only English, but there are plans to add more languages in the future.

Can multiple regulations be analyzed in a single scan?

Yes, customers can select as many regulations as they want when scanning for sensitive data.

Can secrets be discovered in scans ?

Yes, CipherTrust Data Discovery and Classification automatically scans and detects code that will uncover sensitive information that may mistakenly be released by developers.

How do we scan semi-structured data?

CipherTrust DDC scans semi-structured data as unstructured data.

Does the scan engine skip certain files?

It is very common for some scan engines to skip certain files when the file type is unknown. CipherTrust DDC doesn't skip such files, instead it scans them as unstructured files. CipherTrust DDC will skip the files if due to access error it won't be possible to read the content. Those files are going to be considered as inaccessible data objects in the reports.

What is the scan performance?

The scan performance depends on different aspects of the scan such as the number of infotypes in the discover, the size and number of files/ tables, the RAM assigned to the agent(s) involved, and the number of agents assigned to scan the data store(s).

Are the following:

Data Store	Size of data set	Total duration (sec)	Performance (GB/hr)
Linux local storage	Small (917 MB)	340 ~5.6 min	9.5
	Large (1.04 TB)	88759 ~24.66 hs	39.5
MySQL	Small (540 MB)	208.67 ~3.5 min	9
	Medium (9 GB)	2784 ~ 46.4 min	11.3

What does CipherTrust DDC do to reduce false positives and false negatives?

It is important to understand that no solution can eliminate the potential for false positives. CipherTrust DDC uses GLASS, a proprietary technology built in modern era to overcome regex limitations. Ground up design. It leverages modern CPU abilities and provides high performance discovery. It's a powerful feature for complex needs and it scales to petabyte data level.

As a result, CipherTrust DDC solution contains an engine that overcomes many of the legacy functional limitations of Regular Expressions (Regex) for finding data leading to improved accuracy amongst other things. Ultimately the built-in infotypes are equally important in terms of how data is matched and using GLASS, customers can easily evolve their patterns and add additional validation criteria where they store data that is unique to them, but crosses over to common InfoTypes. On top of this, the built-in InfoTypes provides the capability of being configured with high or low precision. High precision means that, CipherTrust DDC will validate the context of the match, searching for specific keywords related with the match. The analysis of the context is going to be done on top of the existing format and algorithm validation. The low precision avoids checking the context. Therefore, if a customer wants to reduce the number of false negatives, they should consider using low precision and take into account that might lead to more false positives. In case they want to reduce the number of false positives, use the high precision selection and bearing in mind that some matches could be missed.

As part of the 250+ InfoTypes which ship within CipherTrust DDC solution and the ability to evolve them per above, it's possible to meet many client specific scenarios or overcome any concerns raised using custom InfoType creation. Also, consider that with the custom infotype capability they can also define the keywords to be used for their specific use case.

Security

Where does the solution store the information it utilizes during operation?

The solution does not collect or store any sensitive data, but it collects the results and type of information an organization manages. This information is stored securely in the Thales Data Platform which is an on-premises Hadoop database, encrypted and hosted by the customer. Thales provides a reference implementation for the customer to use as part of the license agreement.

How does DDC communicate securely with data stores?

CipherTrust Data Discovery and Classification communicates with the data stores through agents. The agents can be installed locally or remotely to the data stores. The agents connect to the data sources using native protocols, e.g., NFS for Unix Share, SMB for Windows Share, HDFS for Hadoop, etc. Each protocol has its own way of protecting data. For example,

- Databases: user and password authentication with SSL/TLS.
- NFS can be secured using host access and file permissions configuration.
- SMB uses user, password and domain authentication.
- Hadoop uses a proprietary protocol.

Customer is responsible for the following:

- Using TLS or plain text
- Hardening of the servers where the agents are deployed

Pricing and Licensing

How is DDC typically packaged and priced?

CipherTrust Data Discovery and Classification is sold as part of the CipherTrust Data Security Platform. The price is based on the amount of data scanned by an organization. An organization can scan one or all of their data stores up to the data amount limit set by the license. CipherTrust Data Discovery and Classification uses data allowance as a way of licensing, providing 1 to 3 year license terms. Once their specified period expires or they use all of their data allowance prior to the expiration of the specified period, whichever occurs first, their access to the solution will cease. If they want to continue using the system, they will need to purchase an additional plan.

How can a customer monitor the data they have consumed?

The customer is able to monitor their data consumption in the admin settings of the CipherTrust Manager console. It displays the total license data amount purchased, the total used to date and the remaining amount available to scan new data. In case the data consumed is higher than the total available, a warning message will be displayed.

How is the data allowance remaining capacity calculated?

The data allowance licensing model is based upon the aggregated maximum size of data ever scanned for each data store that are discovered thus far (the calculation is done per path*). Hence any data changed will not get accounted for any further data allowance consumption, unless the total data size scanned increases, as compared the maximum size ever scanned in the past under that path for the data store.

For example, if in a 4MB file size, only one byte changed, since the overall file-size is not increasing, hence no additional data-allowance capacity will be consumed. Note that, even if the size of a particular file increases, but there are other files whose size decrease to compensate that data size increases in individual files, then the overall data size being scanned is still same or less than the maximum data size scanned in the past and then, the data allowance consumed will not be impacted.

When considering databases, the data allowance consumption model is based upon the aggregated maximum size of data ever scanned from each of the unique data-sources/stores that are scanned to date. For example, if in an Oracle DB with 500GB of data, there is one record updated, re-scanning the database will have no impact on the data allowance consumption. Only does when the database size grows, does it affect the data allowance consumption.

*path configured in the scan, could be a directory, file or table.

What happens at the end of the license period even when unused capacity is available?

After the exhaustion of license period, the customer shall not be able to use the DDC functionality. However, the ability to view old scan reports is still available. It is just new scans are not possible without a valid license in place.

What capacity is provided on renewal of the DDC license?

After renewal, the data allowance capacity stays the same and even if customer did not consume the whole data allowance the previous year, that is not appended to the new one as the validity period finished. Also is important to take into account, the data consumption from the previous year is still consumed in the server after inserting the new license, as it is assumed previous scans and data source will still being scanned.

As an example:

- License 1: 15TB 1 year term

After first year conclude customer only consumed 10TB from 15TB and customer renew license:

- License 2: 15TB, 1 year term

After inserting license 2, the total D.A. will be 15TB and the data consumption will remain 10TB, it is not reset.

The only way to reset the license is doing a fresh installation, but that means losing all the data from the previous year.

THALES

Building a future we can all trust

Contact us

For all office locations and contact information,
please visit cpl.thalesgroup.com/contact-us

cpl.thalesgroup.com