Solution Brief

CipherTrust Data Discovery and Classification Technical Brief

Bringing agility and confidence to your data management

cpl.thalesgroup.com

THALES
Building a future we can all trust

Contents

THALES
Building a future we can all trust

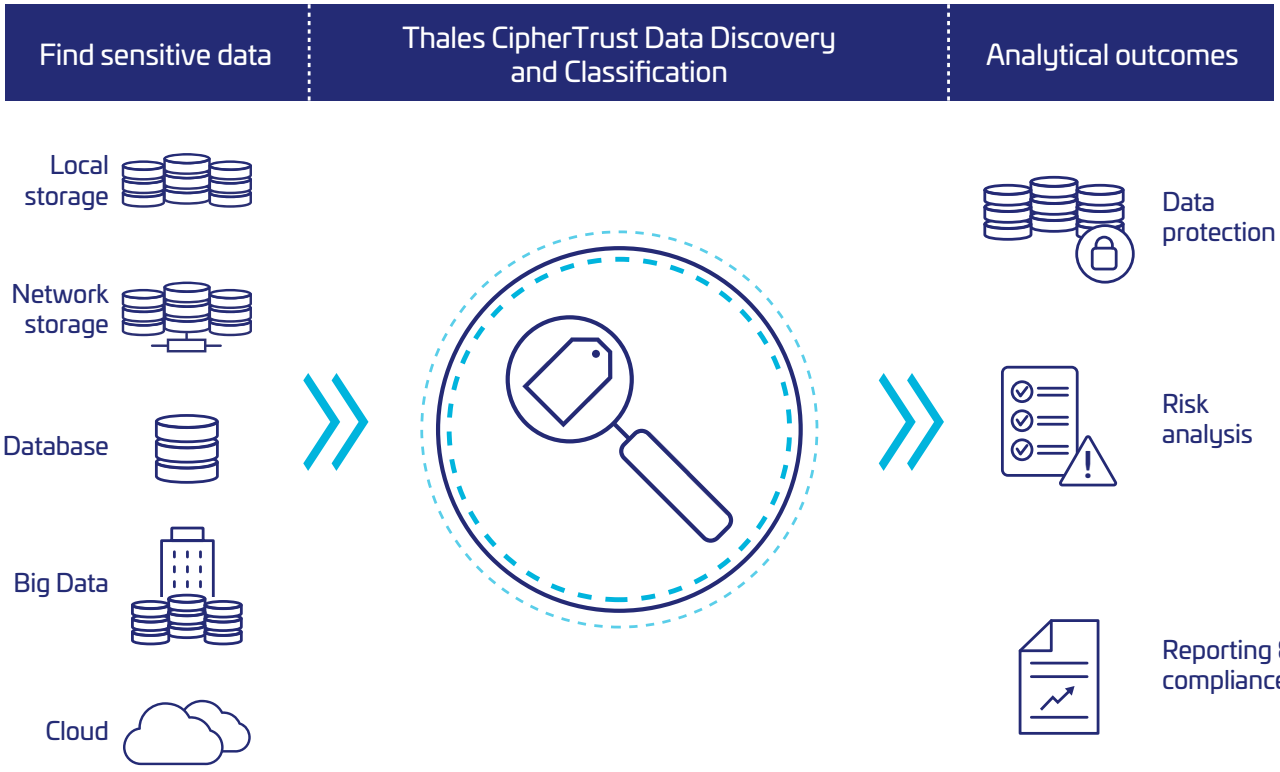
Executive summary

The rapid, often exponential, growth of data year-on-year in organizations like yours makes effective data management an extremely challenging proposition. An increasing switch to remote working raises the bar even higher with many data storage areas or volumes coming into play which are outside your IT team's direct control.

As part of the broader group within your organization responsible for data privacy and security compliance, ensuring no vulnerable areas are overlooked whenimplementing your data protection strategy is of paramount importance. After all, a data breach will inevitably cause severe business disruption in addition to large fines incurred for non-compliance with the seemingly endless stream of new or enhanced data privacy laws and regulations. Failing to prepare properly is definitely not a viable option for you.

A significant weakness often seen in typical data management implementations is the lack of visibility into the precise types of data being held across various local servers, network drives and increasingly cloud storage locations. Many organizations have been breached already, some have encountered near misses and others are migrating large workloads to the cloud without fully understanding the fundamental nature of the data itself and the exposure risks involved – situations for you to avoid.

Historically staff knowledge was sufficient and simple 'off-the-shelf' encryption methods from database vendors would satisfy needs – no longer with the vast data footprint that is widely dispersed and growing by the second. You inevitably need assistance to take control and keep your organization safe.



Thales CipherTrust Data Discovery and Classification is...

A tool that is significantly more effective and efficient than manual methods for data discovery as it scans the whole targets and not just sampling the data. It helps classify all your data, while supporting every mainstream operating system, data storage type, structured and unstructured data you are likely to possess. Existing processes and tools are supplemented, rather than replaced, providing high levels of automation in your search for sensitive data, wherever it resides. Ultimately, it can help you become more agile and support better data management decisions.

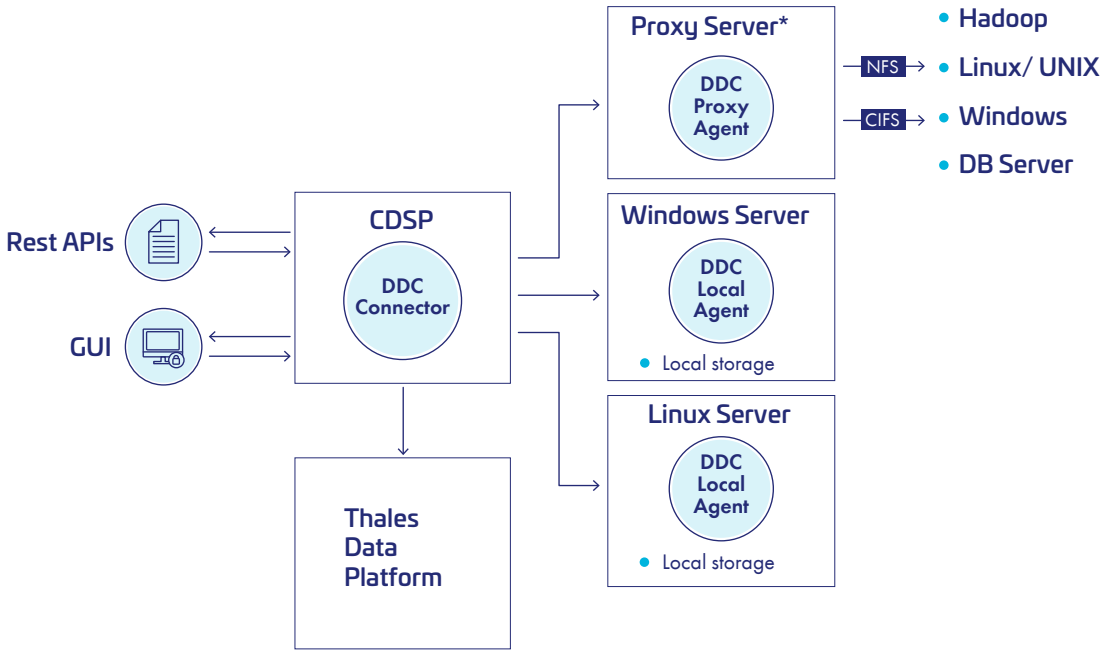
It finds sensitive data that you may not realize even exists in your organization, helping you eliminate threats to your business continuity and reduce your data footprint

It enables you quickly to assess your risks of non-compliance and determine the appropriate protection actions to take

It provides confidence that you are protecting the right assets, including proprietary data or intellectual property, across all your data storage locations including the cloud

The role of CipherTrust Data Discovery and Classification (DDC)

DDC offers a more streamlined, accurate and unattended (or automated) approach in comparison to sensitive data searches using manual methods. DDC stands out from competitive products as it covers additional data storage types, extending the range of structured and unstructured data that can be analyzed while offering higher levels of accuracy by finding sensitive data other tools often miss.



Sensitive items found by DDC are not just restricted to data privacy laws and regulations. Fine tuning using custom infotypes (which you define) facilitates visibility on where data sensitive to your organization is present. By enabling critical assets (such as internal financial data, trade secrets, intellectual property and confidential business plans for example) to be located, their current levels of protection and associated risks of unintended exposure can be better understood. Considerable flexibility comes from running multiple scans simultaneously, each looking for different types of information (if desired) to create a faster and more efficient way to analyze your complete data footprint across a diverse range of storage locations. It is the ability to create multiple classification profiles (customized where needed) that underpins such performance and efficiency benefits that never could be achieved in a comparable timeframe using manual methods or less capable tools.

Discovering the data is only part of the task in hand – classification is equally important, providing deeper understanding of the types of data and their footprint in your organization. DDC does not move or modify your data, merely summarizing what it uncovers during scans in a non-intrusive manner.

You can see **exactly how your data is split** between the personal, financial, medical and national ID categories – you may be surprised that some of this even exists. At least now you will have insight into precisely where it resides to manage the business risk appropriately.

The insights available after scans have been performed provide **important guidance** on what you may need to do next, especially related to protection. You can choose how many scans you wish to incorporate in any given report. Each report improves data visibility, providing detailed insight into:

- The names of the infotypes found together with the number of sensitive data matches
- The risk associated with each of the data objects analyzed during the scan in question
- The protection status so that you can see what it protected and what is not

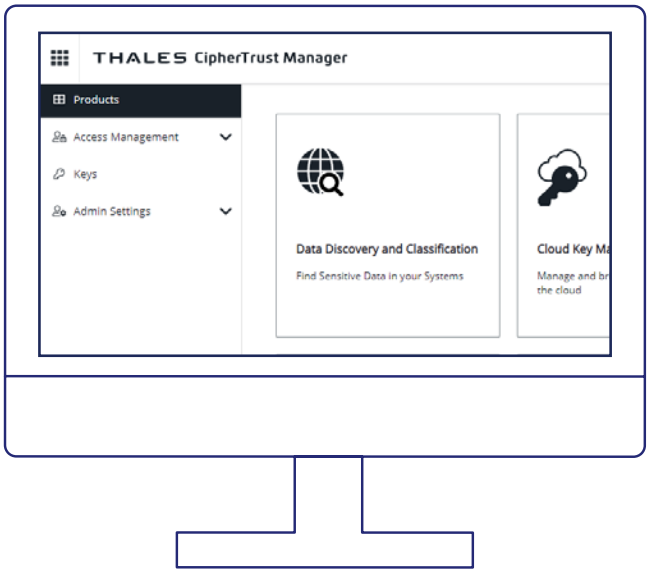
DDC provides options to run scans at any time, over and over again if desired. The tool is intended to be used regularly to keep your insight up to date rather than just being a one-off activity. Since data is always in a state of flux, DDC is ready to convey the latest true representation of what you have and what you might need to do to keep your business safe.

Secrets discovery is a critical component of cybersecurity and data protection. It involves identifying sensitive information, such as API keys, passwords, and social security numbers, that can be inadvertently exposed within systems or applications.

Secrets are difficult to gain visibility to as they reside in so many different places. Developers often scatter them in code repositories, configuration files and personal devices making both the discovery and management of them challenging. Older systems, false positives, and human error make discovering secrets a challenge, especially with cyberattackers constantly coming up with new ways to exploit these vulnerabilities. When secrets such as tokens, API keys, passwords, or usernames are discovered by threat actors, they can be used to break into IT systems. CipherTrust Data Discovery and Classification, using AI, proactively scans code for specific patterns, making developers aware of them before they become security threats. Secrets Discovery is a feature of Thales CipherTrust Data Discovery and Classification product and is the most comprehensive and reliable secrets discovery tool in the marketplace today. It proactively helps stop malicious actors before they gain unauthorized access to your data. Although there are other vendors for secrets discovery, they cannot tell the full story of Discover, Protect, Control like Thales can. The Thales solution is comprehensive, and doesn't require piece mailing solutions, like many competitive solutions do.

How it all fits together

DDC is configured using a graphical user interface (launched from within the CipherTrust Manager console) or via the REST API. It offers a simple and intuitive workflow, enabling you to define data locations for interrogation, specify the types of data you wish to discover and decide when and how often you wish to repeat the scanning process. Ultimately the rich views provide a deep dive into the sensitive data matches, the associated protection status and risk assessment of the data locations in question. The main features are summarized in the sections below, providing a high level insight into the power, flexibility and control available to you.



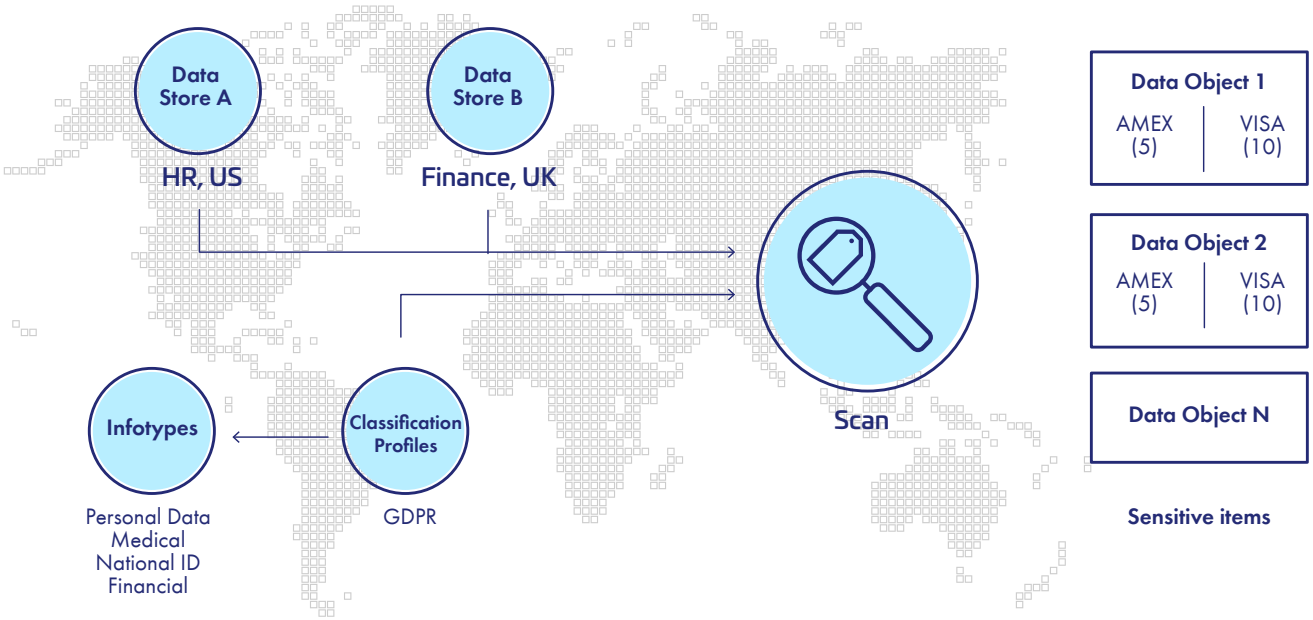
The workflow for DDC is easy to follow via the user interface. **Three main ingredients** help define your environment – **data stores, classification profiles** and **infotypes**.

Typically, you might adopt the following sequence when using DDC for the first time:

Select and configure **data stores** of interest – we recommend concentrating on the most frequently accessed data stores first and a selected path rather than the whole store.

Choose one or more **classification profiles** – create custom classification profiles if necessary. During this process, select one or more **infotypes** of interest - use the custom infotype capability if the prebuilt list is insufficient.

Configure and start the scan – consider just selecting one data store for your first scan to test your configuration.



Data stores		
Local storage	Local storage SharePoint On Prem Exchange Server Local Windows and Linux	Network Storages • Windows Share (CIS/SMB) • Unix File System (NFS)
Network Storage	Windows share (CIS/SMB)	
Database	IBM DB2 Microsoft SQL MongoDS MySQL	Oracle DB PostgreSQL SAP HANA SQL
Big Data	Hadoop clusters	Teradata
Cloud	AWS S3 Buckets Azure Blobs and Table Google Workspace (Gmail and Gdrive)	Azure Table Office 365 (Exchange, SharePoint, & OneDrive) SalesForce

Types of file supported

Databases	Access Dbase	SQLite MSSQL MDF & LDF	**Microsoft Office**	v5 6 95 97 2000	XP 2003 onwards Office Files:Word, Excel, PowerPoint. Access, OutlooK, Other(.pub & .xps)
Images	BMP FAX GIF JPG	PDF (embedded) PNG TIF	**Open Source**	Star Office / Open Office /Libre Office	
Compressed	Bzip2 Gzip (all types)	TAR ZIP (all types)	**Open Standards**	PDF RTF HTML	XML CSV TXT
Microsoft Backup Archive	Microsoft Binary / BKF				

Types of file supported

APA	Australia Privacy Amendment	**HIPAA**	Health Insurance Portability and Accountability Act
APPI	Act on Protection of Personal Information	**KVKK**	Turkish Personal Data Protection Law
CCPA	California Consumer Privacy Act	**LGPD**	General Data Protection Law (Brazil)
GDPR	Financial	**NYDFS**	New York State Department of Financial Services
GDPR	General Data Protection Regulation	**PCI DSS**	Payment Card Industry DataSecurity Standard
GDPR	Healthcare	**SHIELD**	Privacy Shield Framework
GDPR	National ID	**UK-GDPR**	General Data Protection Regulation (UK)
GDPR	Personal Details		

The structured and unstructured data covrage provided by DDC is frequently being updated with new capabilitied - check out the latest specifications here.

Core classification profiles are prebuilt and aligned with major data laws and regulations, both regional and global. You can make changes to these or add totally new profiles if required. Various shortcuts are provided in terms of editing and copying so that you can quickly create and test your own profiles. There is full flexibility in specifying various parameters including sensitivity levels and infotypes.

Types of file supported		
None	**Public**	**Internal**
Private	**Restricted**	

Categorizing sensitive data

DDC is prebuilt with various infotypes, over 250 and counting, covering the vast majority of regional and global data privacy laws and regulations. When a prebuilt classification profile template is selected, the appropriate subset of infotypes is added automatically. You can modify or extend this list by creating your own custom infotypes, allowing rulesto be defined which describe precisely how the scanning engine should look for thedata strings in question. Full access to the groups and categories is available as part of the definition so that data matches can be displayed in the appropriate sections of the scan results.

Infotype categories (prebuilt)		
Financial	Credit/ debit cards	Bank account info
Personal data	Email addresses Login credentials Card number Ethnicity License number Roll number Passport number Date of birth MAC address	Mailing address Telephone number Gender Religion IP address Phone number Name Profanity
Medical	Patient health data	
National ID	Personal identification	
Secrets	Secrets Discovery	

Using agents for discovery

Detailed data analysis is carried out by discovery agents which perform the scanning and report the results back to the DDC Connector for analysis and processing. Two types of agents exist – local and proxy. You would normally install local agents on data store locations (for which you have the appropriate access rights) to ensure the data never leaves the server for security or performance reasons. In contrast, proxy agents are used for network or remote data stores where a separate proxy server hosts the agent.

Type of agent	Pros	Cons
Local	Faster scan Data remains local No need for credential	Longer deployment time More complex management Only supports local data stores
Proxy	Faster to deploy and scale Ability to scan multipledata stores Supports multiple datastore types No resources consumed on target host	Data sent over network to agent prior to scanning Increased network load and data footprint Ideally should be located on same virtual LAN (VLAN)

Common benefits apply to both types of agent:

An **unlimited number** of agents is available for deployment at no additional cost since licensing is based on the volume of new data scanned

Multiple simultaneous sub-scans can be launched to reduce the time required for the overall scan

TLS encryption is available for all communications between data stores and agents to protect sensitive data from eavesdropping

Source data is untouched as agents only access a temporary data copy in internal memory during the scan

No sensitive data is stored during the process as only a summary of the scan results is sent to the DDC Connector

Scans can **still proceed** and record results even when connection to the DDC Connector is lost

With your agents installed, you can then start to think about what and how you wish to scan. We provide a great deal of flexibility in how scans are configured, managed and operated. Important parameters configurable for each scan independently include:

- Name (up to 64 characters)
- Description (up to 250 characters)
- One or more data stores(where to scan)
- One or more classification profiles (what to look for)
- Manual or scheduled operation (when to activate)

On completion of any given scan, it is added to a list of all scans that have been performed. Next, you would use the reporting tool to view detailed results from the scan. Depending on what is found you may make changes to the scan configuration and rerun the scan which is beneficial when working on different data stores or new types of compliance regulations for the first time.

Analyzing scan results

The results from historical and new scans are available for use with the built-in reporting tool. Three main views for scan reports are offered:

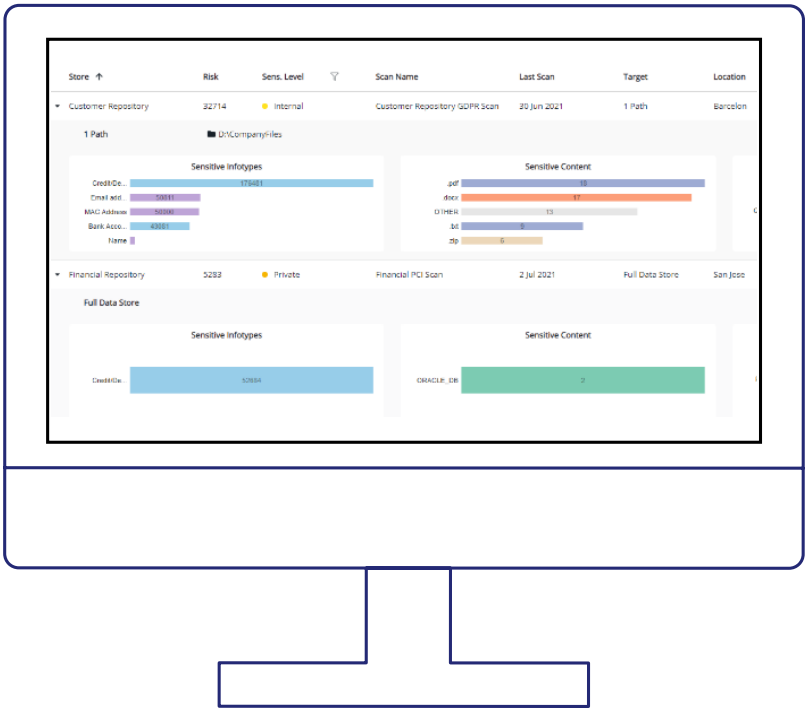
- **Scans**
- **Data stores**
- **Data objects**

Before detailed information is available, a report is generated which aggregates the results from one or more successful scans. Often for logistical, security or performance reasons you may wish to use multiple scans (each looking for different types of regulated data) to analyze your entire data footprint. With our flexible approach you can still do this and fine-tune each scan as required using multiple iterations, before generating a final report which can represent the current status of all your data. It is also easy to understand your organizational posture from a single regulation point of view.

Each tab within the report module offers an insight into slightly different aspects of your data – all however have a common goal in helping you identify quickly any sensitive data at risk.

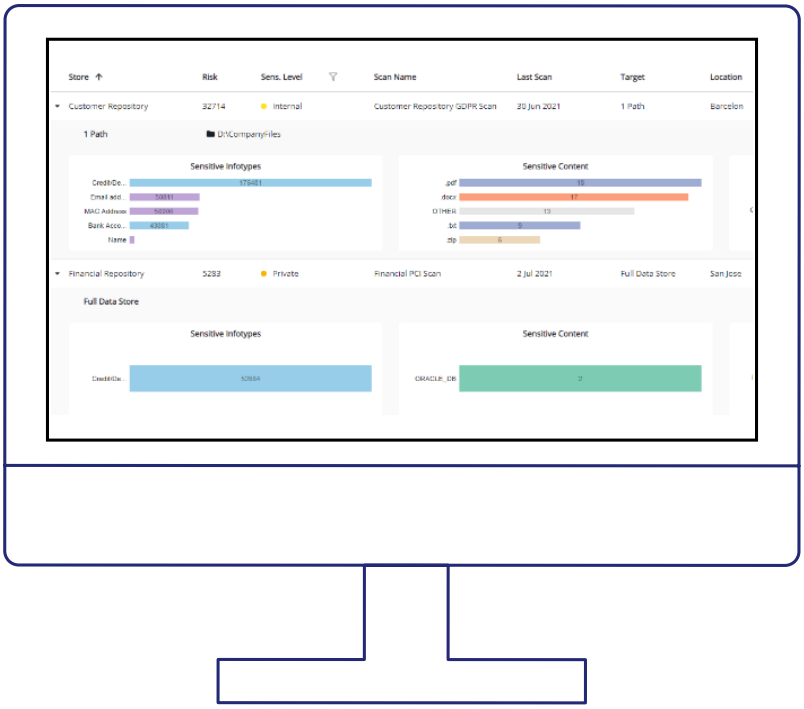
Scans view

Provides a high level, aggregated scan view of the infotypes found (broken down by percentage in each category present) together with three other charts representing the presence of sensitive data objects in terms of category distribution, content types and file types in question.



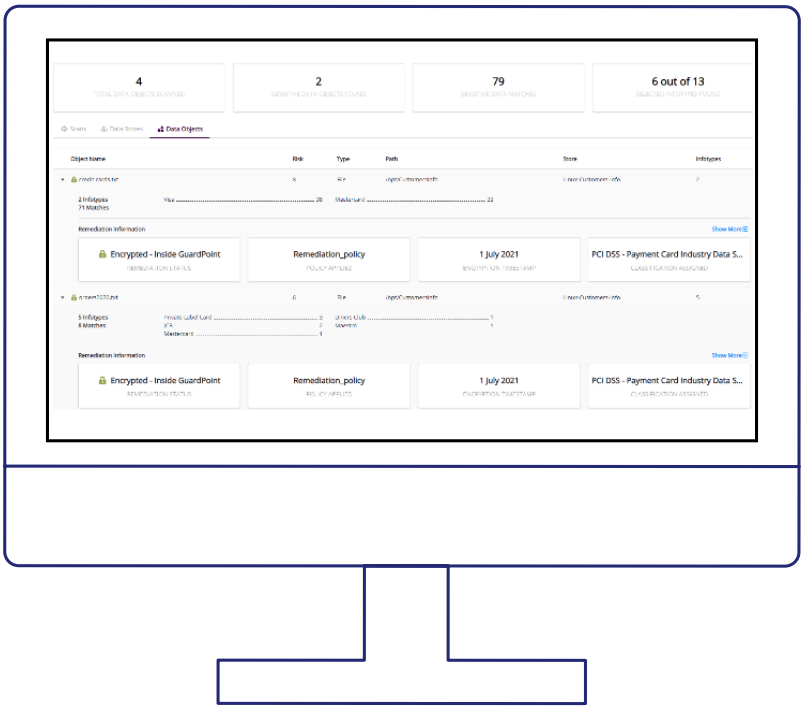
Data stores view

Provides a list of the data stores with details including the risk, sensitivity level, infotypes found and the number of sensitive data objects present for each data store – this provides a rapid means of assessing at a high level whether or not sensitive data is at risk and which data stores are involved.



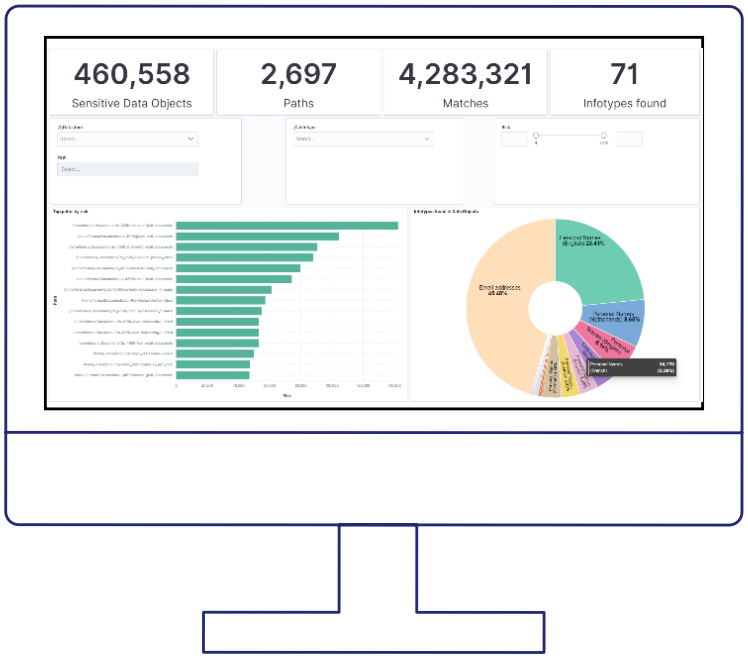
Data objects view

Provides a detailed list of all data objects scanned containing sensitive data, sorted in descending order of risk of exposure. For each sensitive data object in question, youcan see where it is located, the data store in which it resides and the individual names and data matches for all of the associated infotypes found. Remediation status information is also present which provides a deeper insight into your sensitive data, confirming whether or not an encryption policy is active. No sensitive data is stored as part of the reporting process.



Exporting data

In addition to viewing objects using the built-in reporting tool, you can also export data in a NDJSON format for analysis by an external reporting tool such as one from Elastic used in Thales demonstrations



Reasons for our choice of architecture

Single integrated platform

We decided to make DDC an integral part of the Thales CipherTrust Data Security Platform because, in our experience, data discovery is the foundation of any effective data security strategy. This integrated platform approach enables CipherTrust Manager to act as the central management console for various connectors including data discovery, transparent encryption and tokenization. If you are using CTE already, you will be familiar with the console and will benefit from consistency in how common management tasks such as user groups, data policies, access control and logging are implemented.

Accurate pattern matching

Thales performed an extensive investigation to determine the best data discovery technology in the market. After much research and bench testing, the team experienced a significant difference using GLASS TechnologyTM over regex. GLASS is a proprietary market-leading technology fromGround Labs (a Thales Technology Partner) for defining and matching simple and complex patterns of data at scale as opposed to a developer code language not built specifically for purpose. It was designed from the ground up with the express purpose of finding patterns of data in modern data sets across both structured and unstructured data storage scenarios.

High performance discovery

An important goal was to provide a highly efficient and accurate discovery engine. Most data discovery solutions available today use the regular expression (regex) language (originating from the 1950's) to search for patterns inside text strings. This approach requires extensive developer knowledge to implement and the syntax is not based on English grammar. The biggest enemy of regex is performance – as pattern concurrency and complexity increases, the slower the performance achieved. Regex has limits to the patterns it can match accurately, making it far from ideal in the modern world where there are hundreds of different privacy regulations and a broad range of data management issues to address.

Easy customization capabilities

Customization support was another important criteria to supplement the comprehensive prebuilt templates. To address proprietary needs, GLASS data patterns can be written by non-developers using English-style grammar to express how a piece of data or pattern exists. Detailed analysis shows it has overcome all important regex limitations for pattern matching techniques, especially its ability to operate across multiple platform and cloud environments, while being very frugal in terms of CPU overhead. This made GLASS an ideal core component for embedding inside DDC, working efficiently and transparently as part of the CipherTrust platform in identifying and securing critical data.

Flexible licensing model

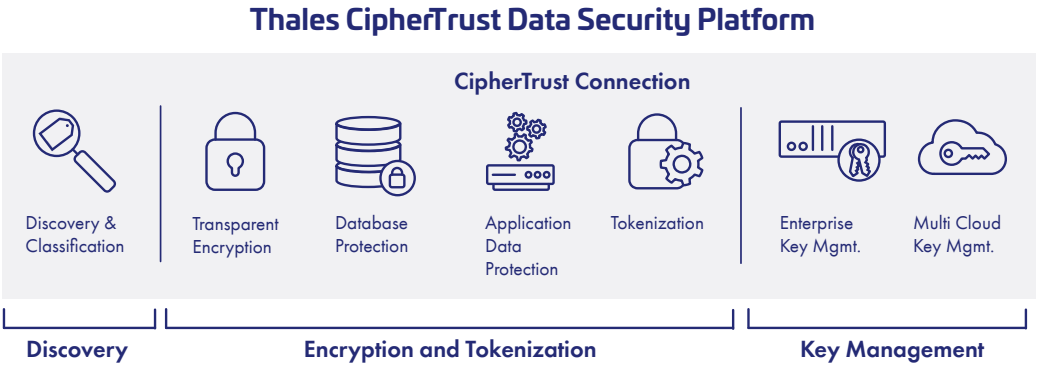
A consumption model based on total data scanned is used to license DDC, providing close alignment with many cloud subscription models and offering an unlimited number of discovery agents without incurring any additional fees. This makes it much easier for all your teams to have access to the data discovery capabilities, with the only real decision to make is how much overall data discovery capacity is required for your organization. Re-scanning existing data does not subtract from your remaining allowance, only new data scanned is counted.

Security first

Naturally as a market leading security company we design and implement secure solutions. DDC is no exception with strong emphasis on ensuring no sensitive data is stored or vulnerable to exposure:

- Ensuring that all data flowing between your data stores and the agents is protected using strong encryption
- Providing granular access controls so that you can tailor the solution to your exact needs

Integrating with other solutions



CipherTrust DataSecurity Platform

DDC is optimized for use as part of the CipherTrust Data Security Platform, an integrated suite of data-centric security products and solutions that unify data discovery, protection and control in one platform. Consequently you can cover all data protection needs using a single platform from one vendor, in this case Thales. This provides you with a full suite of data protection capabilities including key management, user access control, file-level encryption, application-layer encryption, database encryption, format-preserving encryption, tokenization and data masking.

CipherTrust Manager

All components utilize CipherTrust Manager as their management console, delivering central management of encryption keys, granular access control and security policy configuration. One key advantage is that our solution is designed to protect your data wherever it resides, on-premises or in the cloud and whatever storage mechanism is in use, such as files, databases, big data or containers. Ultimately you will end up requiring fewer resources dedicated to data security operations, meet your compliance obligations with greater confidence and reduce the business risk to your organization.

It is also possible to deploy DDC alongside other third party data protection tools, such as those supporting data encryption or tokenization for example. It is non-intrusive, complementing tools you have already, rather than replacing them. However, CipherTrust Manager is still necessary since it provides the sole method to launch the data discovery capability and to manage the licenses required for its operation together with keys required for encrypting sensitive data.

Some of the main reasons Thales customers have been using DDC in conjunction with third party products include:

We often get asked...**'What is the difference between CipherTrust Data Discovery and Classification and other Data Loss Prevention (DLP) solutions?'**

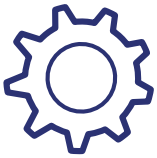
The DLP solutions focus on preventing sensitive data from leaving the organization's perimeter. DDC focuses on data privacy - identifying sensitive data, and getting a clear understanding ofdata and its risk. This enables appropriate steps to be taken to protect data and comply with data privacy and data security regulations.

- Trying to fill gaps in coverage since their existing discovery tool cannot scan all their structured and unstructured data
- Seeking higher accuracy with data matches to eliminate the limitations they are experiencing with regex technology
- Supporting a wider range of operating systems and platforms to interrogate more of their data footprint

Key takeaways

Discover

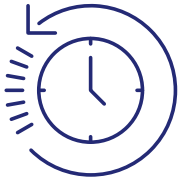
Protect



Control

CipherTrust Data Security Platform

Block unnecessary data access to reduce risk of data breaches



Keep up to date with the evolving regulatory landscape to remain in compliance

Address digital transformation concerns relating to cloud security to support business and technology goals

Your data management and protection strategy can ultimately be improved significantly using CipherTrust Data Discovery and Classification. In conjunction with CipherTrust Manager and CipherTrust Transparent Encryption it provides considerable visibility and insight into your sensitive data while offering quick and effective protection options. It enables you to:

Try our free 90 day evaluation to see for yourself how it can help your organization now

Abbreviations

ACL	Access Control List
CDSP	CipherTrust Data Security Platform
CIFS	Common Internet File System
CM	CipherTrust Manager
CTE	CipherTrust Transparent Encryption
DB	Database
DDC	CipherTrust Data Discovery and Classification
GUI	Graphical User Interface
NAS	Network-attached Storage
REST API	Representational State Transfer Application Programming Interface
TDP	Thales Data Platform
NAS	Network-attached Storage

Glossary

Classification profile	A classification profile uses a list of infotypes to define what kind of sensitive information to search for during a scan
Connector	A connector is a generic term which relates to the different licensable products or components (which includes DDC and CTE) that are managed using the CipherTrust Manager console
CTE agent	A CTE agent is a software component that is used to encrypt data associated with GuardPoints defined using CTE policies
Data match	A data match occurs during a scan when a qualified instance of an infotype actively included in the scope of the search is found
Data object	A file or database table located inside a data store is known as a data object
Data store	A data store is the entity where the data is actually stored, with DDC supporting various types - local, network, database, Big Data and cloud
DDC agent	A DDC agent is a software component that is used to scan a data store for specific types of data defined using infotypes in the associated
Discovery	classification profile
False negatives	Data discovery is the process of mapping an organization's data assets including their locations, types and sensitivity levels
False positives	False negatives are where the discovery process fails to identify one or more valid instances of sensitive data
GuardPoint	A GuardPoint specifies the list of folders or paths to be protected –access to files and encryption of files under the GuardPoint is controlled by security policies
Infotype	An infotype is used to categorize specific data (such as passport numbers or email addresses) to look for during a discovery scan, forming an integral component in the definition of a classification profile
Policy	A policy is a collection of rules that govern data access and encryption
Remediation	Remediation refers to the process of protecting (or securing) data that has been identified as vulnerable as the result of a scan – typical remediation methods include encryption, tokenization, data masking and access control
Resource set	A resource set refers to the files or directories to which the policy will apply together with the associated governing key rules
Risk	A risk is the presence of a sensitive data object in a data store and is directly related to the data matches found in the data object or data store
Scan	A scan is part of the discovery process that is used to search for sensitive data within data stores using criteria defined in classification profiles
Sensitive data object	A data object that contains any data match is known as a sensitive data object

Sensitivity level	The sensitivity level is a mandatory parameter (when defining data stores and classification profiles) relating to the degree of potential vulnerability of any given data object if exposed
Structured data	Structured data is highly organized and easily understood by machine language – examples include names, dates, addresses, credit card numbers, stock information, geolocation and more
Tag	A tag helps group data together and can be specified when creating data stores and classification profiles
Thales Data Platform	The Thales Data Platform (TDP) is a Big Data platform based on Hadoop technology which is used by DDC for various tasks including storage ofscan results
Unstructured data	Unstructured data is information that either does not have a pre-defined data model (or schema) or is not organized in a pre-defined manner, making it unsuitable for storage in a relational database

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