



Classify Data: SITs, EDM, Fingerprints, Classifiers

SC-401 domain 1, Implement information protection (part 1) | Sensitive information types, confidence and proximity, exact data match, document fingerprinting, trainable classifiers, content and activity explorer

SC-401 Domain 1 Data classification Purview

Know your data: four ways to classify

Data classification in Purview

×1

Overview

Purview reads content and decides what it is: a card number, a contract, a patient file.

×10

Close up

Four classifier families, from rule-based to learned:

- sensitive information types: patterns, keywords, checksums
- exact data match: your own values, hashed
- document fingerprinting: derived from a template form
- trainable classifiers: ML on document kinds

×100

Under the microscope

Every later control reuses the result:

- auto-labelling conditions
- DLP rule conditions
- retention label auto-apply
- insider risk indicators (sensitive content)
- DSPM for AI reports

Category	Detects	Built from	Best for
SIT (pattern)	formatted values	regex, keywords, checksum function	card, IBAN, passport, national ID, keys
EDM	your exact records	hashed table + schema	customer or patient IDs, low false positives
Fingerprint	documents from a form	one template file	patent, HR, tax forms
Trainable classifier	kinds of documents	pre-trained, or 50-500 seed samples	contracts, resumes, source code, threats
Named entity	names, addresses, medical terms	dictionaries + ML	GDPR personal data in prose

Built-in first

Purview ships 300+ built-in SITs (countries' IDs, finance, health, credentials such as A11 credential types and Azure storage keys). You cannot edit a built-in SIT: **copy** it, then change the copy (Data classification → Classifiers → Sensitive info types → Copy). Test any SIT on a sample file with the **Test** button or `Test-DataClassification`.

Anatomy of a SIT pattern

Part	What it does	Example (credit card)
Primary element	must match: regex, keyword list, keyword dictionary or built-in function	Func_credit_card: 16-digit formats with the Luhn checksum
Supporting elements	raise confidence when found within the proximity window	keywords 'Visa', 'card verification', an expiry date
Proximity	character window around the primary element (default 300 characters)	keyword within 300 characters
Confidence level	high 85, medium 75, low 65 per pattern	high: number + keyword or date; medium: number with valid checksum
Additional checks	exclude specific values, include or exclude prefixes and suffixes, exclude duplicate characters	exclude test numbers 4111 1111 1111 1111
Instance count	set in the DLP or auto-labelling rule, not in the SIT	1-9 low volume, 10+ high volume

\bTD-20\d{2}-\d{6}\b

\b	a word boundary
TD-20	the text "TD-20"
\d{2}	a digit 0-9, exactly 2 times
-	the character "-"
\d{6}	a digit 0-9, exactly 6 times
\b	a word boundary

- ✓ "Contract TD-2026-004512 signed" matches: "TD-2026-004512"
- ✗ "TD-26-004512" does not match
- ✗ "ref xTD-2026-004512" does not match
- ✓ "TD-2031-990001 renewal" matches: "TD-2031-990001"

A custom SIT primary element: Tadata contract numbers like TD-2026-004512

Conditions	R1	R2	R3	R4
Value has a fixed format (regex or checksum)	Y	--	--	N
Exact values exist in your database	--	Y	--	N
Documents come from one standard form	--	--	Y	N
Only the kind of document is known	N	N	N	Y
Actions				
Custom or built-in SIT	X			
EDM SIT		X		
Document fingerprint			X	
Trainable classifier				X
Notes	Add supporting keywords to cut false positives	Use a SIT as the EDM primary element to find candidates	Upload the blank form	Pre-trained first; custom needs 50-500 positive seeds

Which classifier for which need

📄 **Worked example: confidence levels and instance counts in a rule**

A spreadsheet holds N = 12 card numbers: H = 8 match at high confidence (checksum + 'Visa' within 300 characters) and M = 4 only at medium (checksum, no keyword). Rule A requires min count 10 at high confidence: it sees H = 8 < 10 → no match. Rule B requires min count 10 at medium confidence: a medium threshold also counts high matches, H + M = 12 ≥ 10 → match. Lesson: lowering the confidence raises recall; raising the count lowers false positives.

📖 **Exact data match and fingerprints**

- 1 **Prepare the table** CSV/TSV export of the sensitive records (patient ID, name, birth date); one row per person
- 2 **Create the EDM schema** portal wizard (or XML): columns, which are searchable (the values to look for), case and punctuation handling
- 3 **Create the EDM SIT** primary element = a searchable column mapped to an existing SIT that finds candidates; supporting elements = other columns within proximity
- 4 **Hash and upload** EDM Upload Agent on a secure machine, member of the EDM_DataUploaders group; data is salted and hashed, the cleartext never leaves you
- 5 **Use it** in DLP, auto-labelling and IRM conditions like any SIT; schedule refreshes when the source changes

EDM setup, in order

Command	Purpose
EdmUploadAgent.exe /Authorize	sign in the agent as an EDM_DataUploaders member
EdmUploadAgent.exe /CreateHash /DataFile pat.csv /HashLocation C:\h /Schema s.xml	hash locally (separate machine from upload when policy requires)
EdmUploadAgent.exe /UploadHash /DataStoreName patients /HashFile C:\h\pat.EdmHash	upload the salted hashes
EdmUploadAgent.exe /GetSession /DataStoreName patients	check the indexing status of the upload
Test-DataClassification -TextToClassify 'Card 4539 1488 0343 6467 Visa'	show which SITs and confidence a text triggers
Get-DlpSensitiveInformationType ? Publisher -ne 'Microsoft Corporation'	list custom SITs

Document fingerprinting	Trainable classifiers
upload a blank template (Word, PDF, Excel); Purview stores word patterns, not the file	pre-trained: agreements, financial statements, resumes, source code, harassment, profanity, threat and more
partial or exact matching thresholds for filled-in copies	custom: 50-500 positive seed samples in a SharePoint folder, then review test predictions and publish
used in DLP for Exchange, SharePoint, OneDrive, Teams and devices	processing takes hours to days; retrain with match / not match feedback
PowerShell: New-DlpFingerprint then New-DlpSensitiveInformationType -Fingerprints	usable in auto-labelling (Office and service-side), DLP, retention auto-apply, communication compliance

👁 **See what you have: explorers**

Content explorer	Activity explorer
what is stored: items at rest with a SIT, label or classifier	what happened: label applied, changed, removed; DLP match; file copied, printed, uploaded
by location: Exchange, SharePoint, OneDrive, Teams, devices	last 30 days of events, filters by activity, user, label, location
List Viewer sees names; Content Viewer opens the item	built from the unified audit log: audit must be on item
counts can lag behind changes (index refresh)	export to CSV for a longer record

MYTH	FACT
✗ A medium-confidence rule ignores high-confidence matches	✓ Lower thresholds include higher levels Medium counts high and medium; low counts all three.
✗ EDM uploads my customer table to Microsoft	✓ Only salted hashes are uploaded The cleartext stays on your machine; the hashes are matched against hashed candidates.
✗ Built-in SITs can be tuned in place	✓ Copy them first Built-in SITs are read-only; the copy becomes a custom SIT you can edit.
✗ Trainable classifiers need a regex	✓ They learn from example documents Use them when meaning, not format, defines the data.

⚠ False positives cost trust

A raw 9-digit regex for a national ID will match phone numbers and order IDs. Always add supporting keywords, a checksum function when one exists, and exclusions; then pilot DLP in simulation and read the matches in activity explorer before enforcing.

Q1 Default proximity window of a SIT pattern?

A 300 characters around the primary element.

Q2 Numeric values of high, medium and low confidence?

A 85, 75 and 65.

Q3 Your customer IDs have no fixed format but sit in a CRM table: which classifier?

A Exact data match (EDM) built from a hashed CRM export.

Q4 Which group must the EDM upload account belong to?

A EDM_DataUploaders.

Q5 Detect filled-in copies of the company's patent application form?

A Document fingerprinting from the blank form.

Q6 How many seed samples does a custom trainable classifier need?

A Between 50 and 500 positive samples, stored in a SharePoint folder.

Q7 Who can open the actual content of a classified file in content explorer?

A Members of Content Explorer Content Viewer (List Viewer sees only the list).

Q8 How far back does activity explorer show events?

A 30 days; export or use the audit log for longer.

Q9 Which cmdlet tests text against SITs?

A Test-DataClassification -TextToClassify '...'.
`Test-DataClassification -TextToClassify '...'`

Key takeaways

01

Format → SIT, records → EDM, form → fingerprint, meaning → classifier

pick by the shape of the data

02

Confidence and count are tuning knobs

confidence in the SIT, instance count in the rule

03

Copy, then customise

built-in SITs are read-only

04

Hashes only for EDM

EDM Upload Agent, EDM_DataUploaders

05

Explorers prove it

content explorer for what is there, activity explorer for what happened