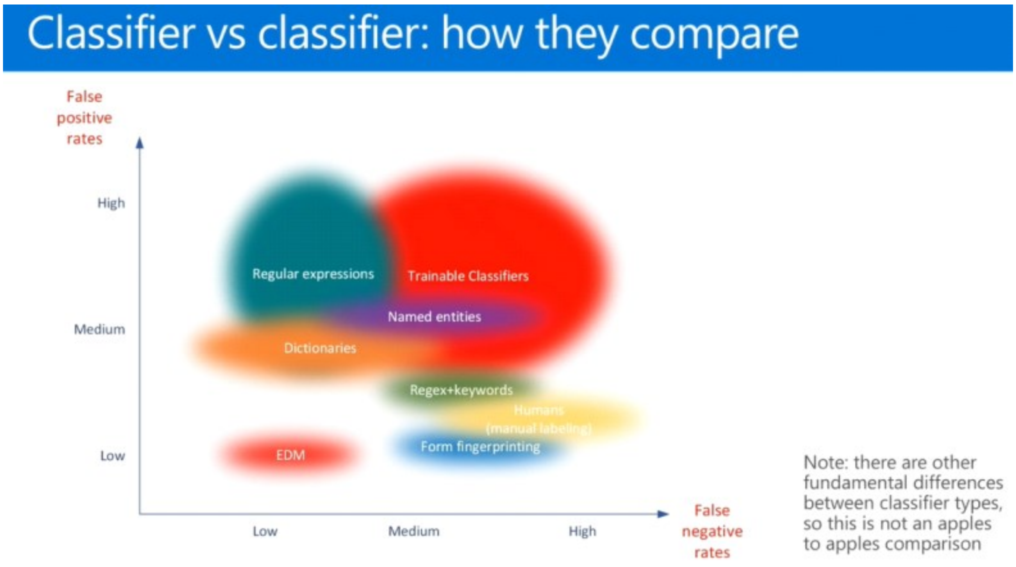
Webinaire Exact Data Match

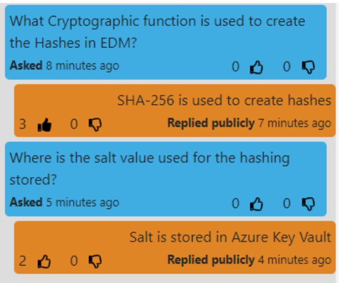
mercredi 18 mai 2022 17:15

[18:05] Robin Baldwin (OLIVE & GOOSE LLC)
Andrew Wadler (External)
is there a link to the prior webinar?
https://mipc.eventbuilder.com/event/61024
like 2
Configuring Exact Data Matching for Accurate Data Classification
When trying to protect sensitive data about known subjects such as customer or employee PII, Exact Data Matching - thanks to its ability to precisely target the right data with almost zero false po...



Classifier vs classifier: how they compare

Note: there are other fundamental differences between classifier types, so this is not an apples to apples comparison

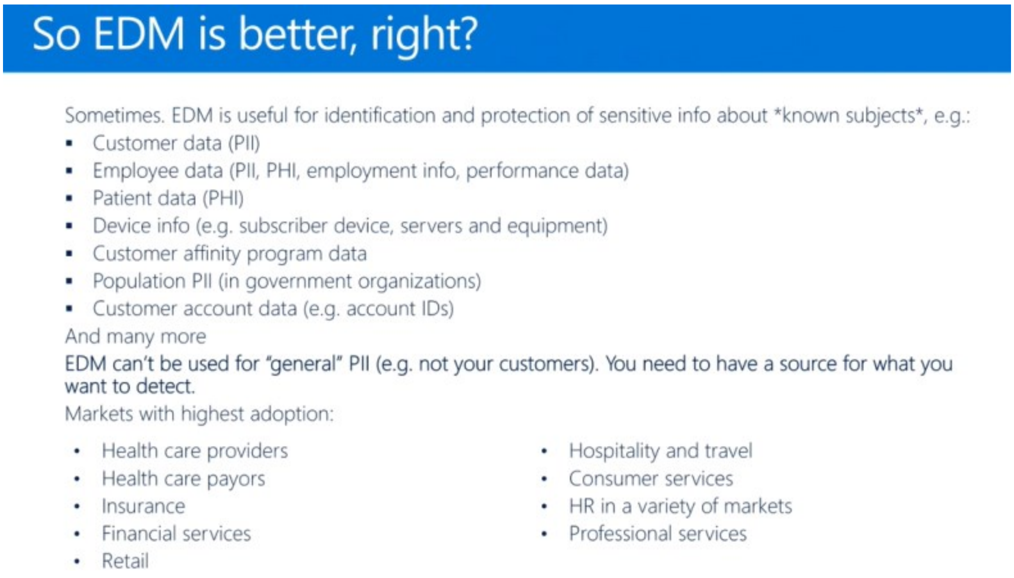


What Cryptographic function is used to create the Hashes in EDM?
Asked 8 minutes ago 0 👍 0 👎

SHA-256 is used to create hashes
3 👍 0 👎 **Replied publicly** 7 minutes ago

Where is the salt value used for the hashing stored?
Asked 5 minutes ago 0 👍 0 👎

Salt is stored in Azure Key Vault
2 👍 0 👎 **Replied publicly** 4 minutes ago



So EDM is better, right?

Sometimes. EDM is useful for identification and protection of sensitive info about *known subjects*, e.g.:

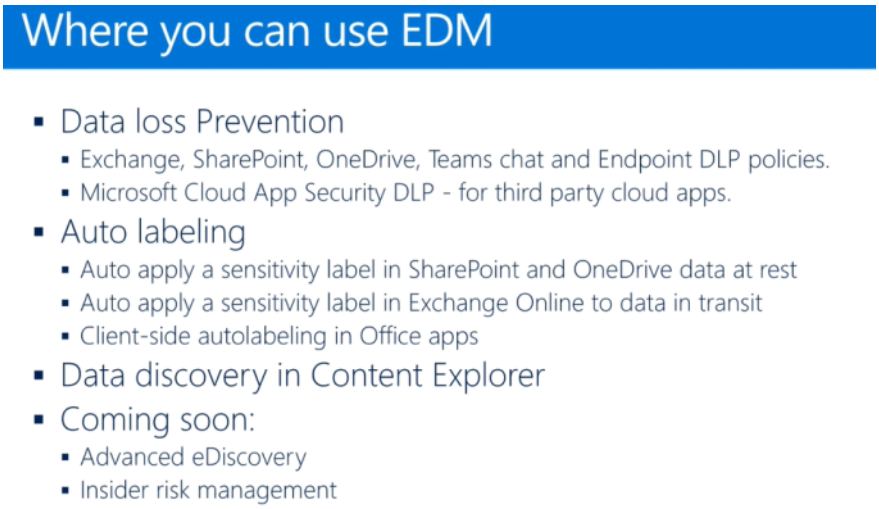
- Customer data (PII)
- Employee data (PII, PHI, employment info, performance data)
- Patient data (PHI)
- Device info (e.g. subscriber device, servers and equipment)
- Customer affinity program data
- Population PII (in government organizations)
- Customer account data (e.g. account IDs)

And many more

EDM can't be used for "general" PII (e.g. not your customers). You need to have a source for what you want to detect.

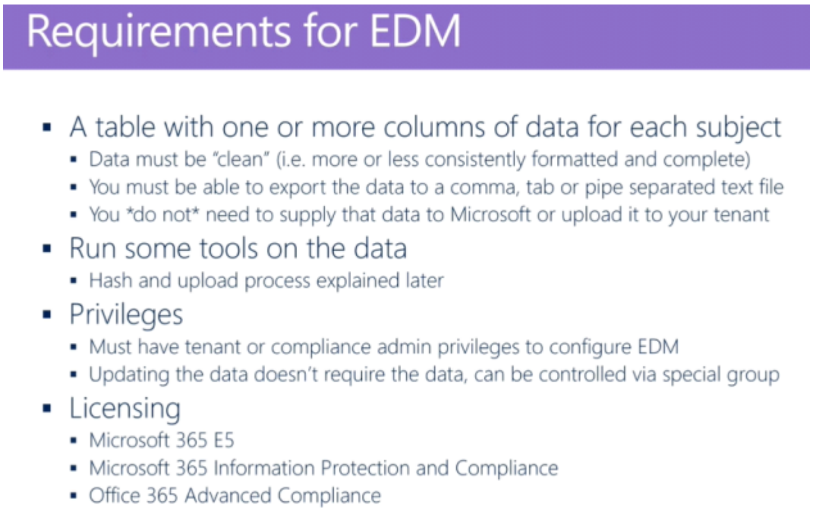
Markets with highest adoption:

- Health care providers
- Health care payors
- Insurance
- Financial services
- Retail
- Hospitality and travel
- Consumer services
- HR in a variety of markets
- Professional services



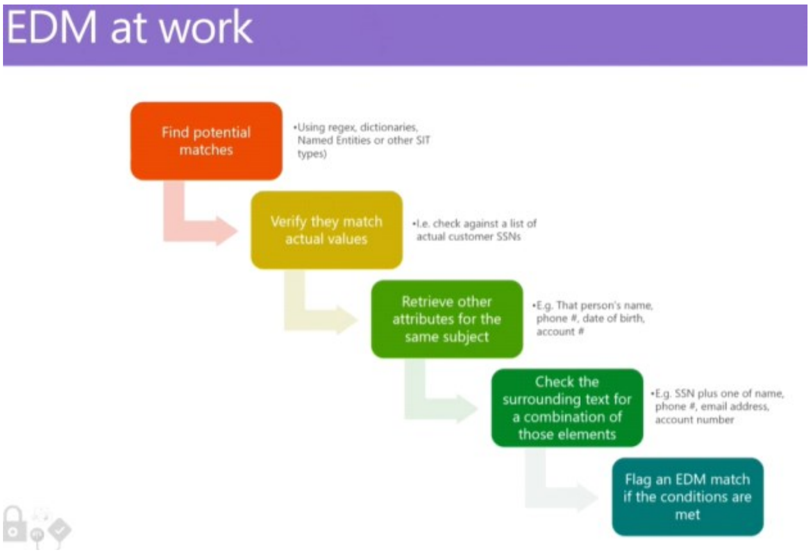
Where you can use EDM

- Data loss Prevention
  - Exchange, SharePoint, OneDrive, Teams chat and Endpoint DLP policies.
  - Microsoft Cloud App Security DLP - for third party cloud apps.
- Auto labeling
  - Auto apply a sensitivity label in SharePoint and OneDrive data at rest
  - Auto apply a sensitivity label in Exchange Online to data in transit
  - Client-side autolabeling in Office apps
- Data discovery in Content Explorer
- Coming soon:
  - Advanced eDiscovery
  - Insider risk management



Requirements for EDM

- A table with one or more columns of data for each subject
  - Data must be "clean" (i.e. more or less consistently formatted and complete)
  - You must be able to export the data to a comma, tab or pipe separated text file
  - You *do not* need to supply that data to Microsoft or upload it to your tenant
- Run some tools on the data
  - Hash and upload process explained later
- Privileges
  - Must have tenant or compliance admin privileges to configure EDM
  - Updating the data doesn't require the data, can be controlled via special group
- Licensing
  - Microsoft 365 E5
  - Microsoft 365 Information Protection and Compliance
  - Office 365 Advanced Compliance



EDM at work

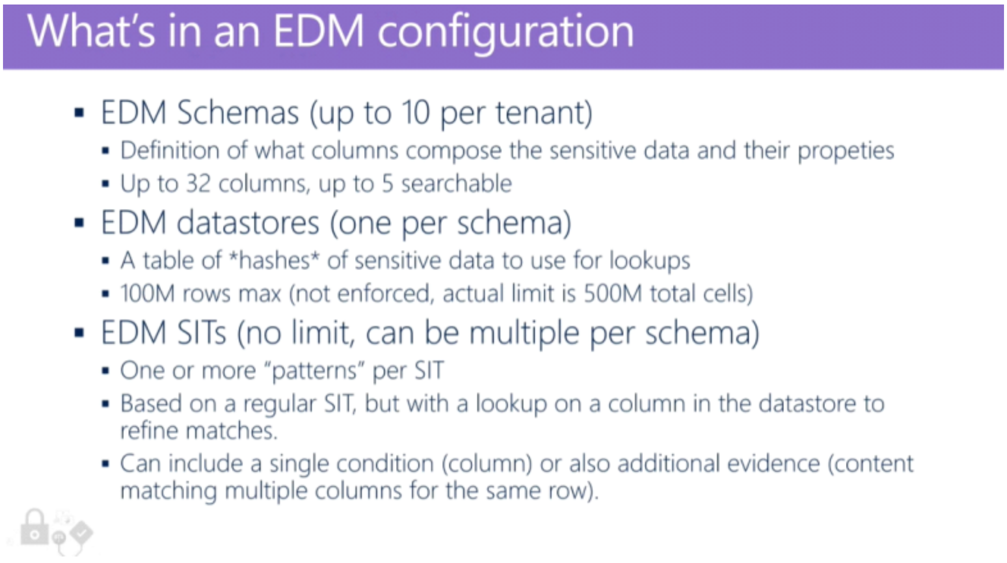
Find potential matches
•Using regex, dictionaries, Named Entities or other SIT types)

Verify they match actual values
•I.e. check against a list of actual customer SSNs

Retrieve other attributes for the same subject
•E.g. That person's name, phone #, date of birth, account #

Check the surrounding text for a combination of those elements
•E.g. SSN plus one of name, phone #, email address, account number

Flag an EDM match if the conditions are met



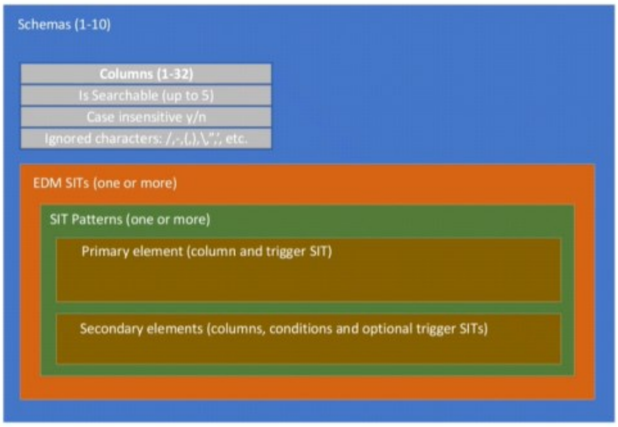
What's in an EDM configuration

- EDM Schemas (up to 10 per tenant)
  - Definition of what columns compose the sensitive data and their propeties
  - Up to 32 columns, up to 5 searchable
- EDM datastores (one per schema)
  - A table of *hashes* of sensitive data to use for lookups
  - 100M rows max (not enforced, actual limit is 500M total cells)
- EDM SITs (no limit, can be multiple per schema)
  - One or more "patterns" per SIT
  - Based on a regular SIT, but with a lookup on a column in the datastore to refine matches.
  - Can include a single condition (column) or also additional evidence (content matching multiple columns for the same row).

Detour: why do we have to identify "potential" matches?

- Can't we just check everything for an exact match?
  - We could, but it would be computationally impractical and slow.
  - A match can be a word or number, multiple words, part of a word, etc.. Each document contains $(n^2+n)/2$ sequences of strings inside (not counting ignored delimiters or casing)
  - If your company has 50,000 employees producing 100 pieces of content (email or document) per second, each with 500 words, that is equivalent to three quintillion strings to check per day.
  - If your table has 100 million rows and ten columns to check against each… you get the idea.
- But can't you optimize it? Not all sequences make sense!
  - That's what we did: you tell us (via a SIT) what's a meaningful string to check.

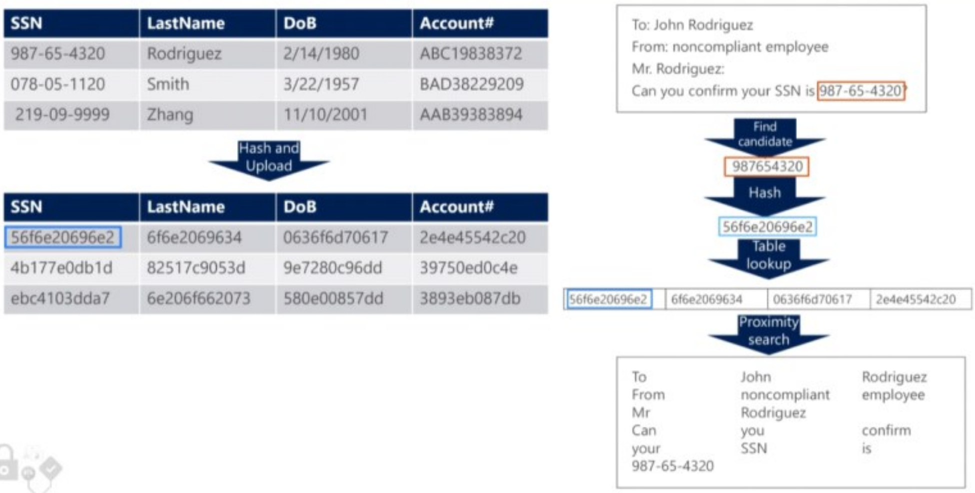
Anatomy of an EDM SIT


Schemas (1-10)
Columns (1-32)
Is Searchable (up to 5)
Case insensitive y/n
Ignored characters: /,-,(,),\,",', etc.
EDM SITs (one or more)
SIT Patterns (one or more)
Primary element (column and trigger SIT)
Secondary elements (columns, conditions and optional trigger SITs)

About the sensitive info table

- For EDM to work we *must* be able to check candidate matches against your data.
- But we don't need your data!!!!
- We can use this little trick called hashing
  - A hash is a non-reversible but quasi unique transformation of a string:
  - E.g.:
  - The cat in the hat => 7a652063617374e0696e207468f52f68617
  - The cat in the hut => bb2206c6d20cd04bb46t6161ae206c21db
  - 7a65206361737420696e207468652068617 ≠> The cat in the hat
- No other plausible text matches those hashes
- Only way to find out the original value is to hash all possible values and compare
- A hash can be "salted" by adding a fixed value to each string before hashing, to make the transformation unique to the customer

How hashing is used in EDM



SSN	LastName	DoB	Account#
987-65-4320	Rodriguez	2/14/1980	ABC19838372
078-05-1120	Smith	3/22/1957	BAD38229209
219-09-9999	Zhang	11/10/2001	AAB39383894

SSN	LastName	DoB	Account#
56f6e20696e2	6f6e2069634	0636f6d70617	2e4e45542c20
4b177e0db1d	82517c9053d	9e7280c96dd	39750ed0c4e
ebc4103dda7	6e206f662073	580e00857dd	3893eb087db

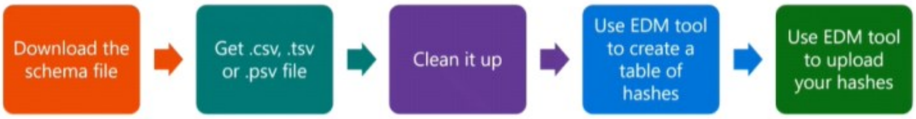
To: John Rodriguez
From: noncompliant employee
Mr. Rodriguez:
Can you confirm your SSN is 987-65-4320?

56f6e20696e2	6f6e2069634	0636f6d70617	2e4e45542c20

To	John	Rodriguez
From	noncompliant	employee
Mr	Rodriguez	
Can	you	confirm
your	SSN	is
987-65-4320		

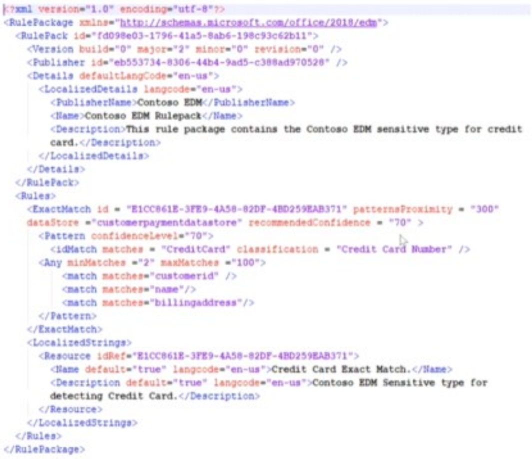
Deploying EDM (short version)

- Step 1: Create your schema
  - From compliance center or via PowerShell
- Step 2: Define your sensitive info types
  - Can be done after step 3 or more likely in parallel
- Step 3: Hash and upload your data


Download the schema file → Get .csv, .tsv or .psv file → Clean it up → Use EDM tool to create a table of hashes → Use EDM tool to upload your hashes

- Step 4: Profit!

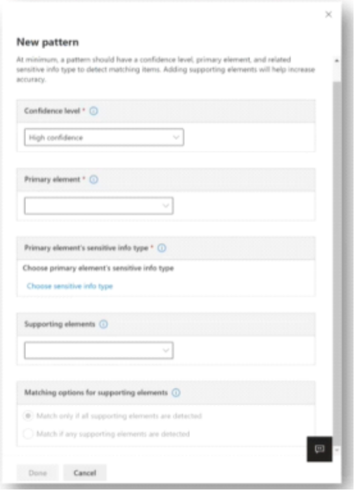
Define EDM Sensitive Information type (the hard way)


```
<?xml version="1.0" encoding="utf-8"?>
<RulePackage xmlns="http://schemas.microsoft.com/office/2018/edm">
  <RulePack id="fd098e03-1796-41a5-8ab6-198c93c62b11">
    <Version build="0" major="2" minor="0" revision="0" />
    <Publisher id="eb553734-8306-44b4-9ad5-c388ad970528" />
    <Details defaultLangCode="en-us">
      <LocalizedDetails langcode="en-us">
        <PublisherName>Contoso EDM</PublisherName>
        <Name>Contoso EDM Rulepack</Name>
        <Description>This rule package contains the Contoso EDM sensitive type for credit
        card.</Description>
      </LocalizedDetails>
    </Details>
  </RulePack>
  <Rules>
    <ExactMatch id = "E1CC861E-3FE9-4A58-82DF-4BD259EAB371" patternsProximity = "300"
    dataStore ="customerpaymentdatastore" recommendedConfidence = "70" >
      <Pattern confidenceLevel="70">
        <idMatch matches = "CreditCard" classification = "Credit Card Number" />
      <Any minMatches ="2" maxMatches ="100">
          <match matches="customerid" />
          <match matches="name"/>
          <match matches="billingaddress"/>
      </Pattern>
    </ExactMatch>
    <LocalizedStrings>
      <Resource idRef="E1CC861E-3FE9-4A58-82DF-4BD259EAB371">
        <Name default="true" langcode="en-us">Credit Card Exact Match.</Name>
        <Description default="true" langcode="en-us">Contoso EDM Sensitive type for
        detecting Credit Card.</Description>
      </Resource>
    </LocalizedStrings>
  </Rules>
</RulePackage>
```

Sample xml for EDM Sensitive type

Define an EDM SIT (the slightly easier way)

- In the Exact Data Match section of the Compliance Center, select EDM Sensitive Types
- Create a new EDM type
- Select your schema
- Create one or more patterns
  - Select primary element (column)
  - Select a SIT that describes it
  - Select columns to use as secondary element
  - Define matching rules (e.g. one of n, all, etc.)


New pattern

At minimum, a pattern should have a confidence level, primary element, and related sensitive info type to detect matching items. Adding supporting elements will help increase accuracy.

Confidence level *: High confidence

Primary element *:

Primary element's sensitive info type *: Choose primary element's sensitive info type

Supporting elements:

Matching options for supporting elements

● Match only if all supporting elements are detected

○ Match if any supporting elements are detected

Hash and upload your sensitive data

EDM Upload Agent Purpose:

To one-way hash the data to have a file to upload

To Upload file with hashes to the service – where it is stored and ready for lookups

Uploading the file also upload the (automatically generated or manually entered) salt used for hashing

Set up the security group and user account

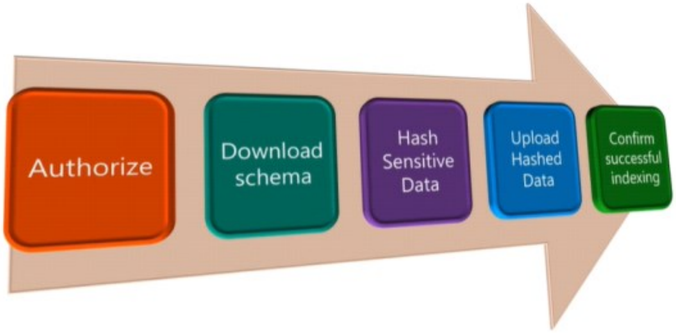
- As a global administrator, go to the admin center create a security group: EDM_DataUploaders.
- Add one or more users to the EDM_DataUploaders security group.

- Set up the EDM Upload Agent
  - Download EDM upload agent
  - https://go.microsoft.com/fwlink/?linkid=2088639
  - Agent trace logs located:

  C:\Program Files\Microsoft\EdmUploadAgent\TraceLogs

  Tip: DO NOT install in default folder (program files), use a custom folder, so you do not need admin privileges in the machine to use it.

Agent Workflow



Authorize

Download schema

Hash Sensitive Data

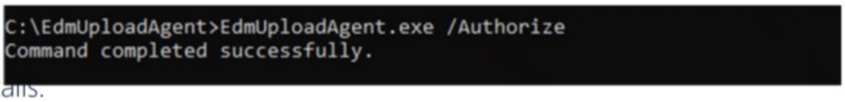
Upload Hashed Data

Confirm successful indexing

Authorize Agent

`EdmUploadAgent.exe /Authorize`

Example:



```
C:\EdmUploadAgent>EdmUploadAgent.exe /Authorize
Command completed successfully.
```

Details:

- This will prompt for user credentials to authorize the EDM upload agent to act on behalf of the user.
- It is recommended to create a separate dedicated user with minimal privileges which can be used for EDM Upload agent.*
- Authorization must be done every 30 days (depending on your tenant's AAD auth token configuration)
- Re-run the Authorize command, if any other command fails with authorization errors.
- Please note: there's an Authorize.ps1 script you can use to pass credentials interactively or script so you can pass them as a SecureString

Hash and upload Sensitive Data

```
EdmUploadAgent.exe /CreateHash /DataStoreName <DataStoreName> /DataFile <DataFilePath> /HashLocation <HashedFileLocation>
EdmUploadAgent.exe /UploadHash /DataStoreName <DataStoreName> /HashFile <HashedSourceFilePath>
```

Example:

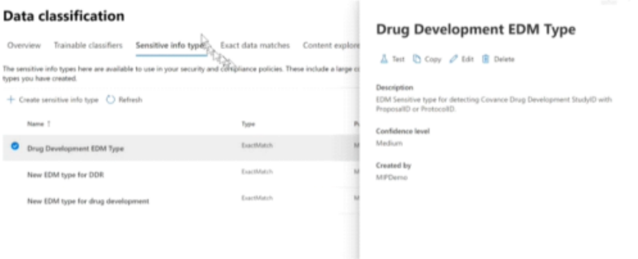
```
C:\EdmUploadAgent>EdmUploadAgent.exe /CreateHash /DataStoreName patient /DataFile C:\BugBash\EDM\Patient.csv
 /HashLocation C:\BugBash\EDM
Command completed successfully.

C:\EdmUploadAgent>EdmUploadAgent.exe /UploadHash /DataStoreName patient /HashFile C:\BugBash\EDM\Patient.EdmHash
Command completed successfully.
```

Details:

- DataStoreName: The name of the data store whose schema has already been defined. Hint: same name as the schema.
- DataFile: Provide the full path to the data file.
- HashLocation: Provide the path to the folder where the hash file should be created.
- The naming format of the hash file created is "datafilename.EdmHash".

Testing EDM Sensitive Info Type

- Classifications, Sensitive info type
- Select EDM type
- Upload file to test



Data classification

The sensitive info types here are available to use in your security and compliance policies. These include a large c... types you have created.

Name ↑	Type
Drug Development EDM Type	ExactMatch
New EDM type for DDR	ExactMatch
New EDM type for drug development	ExactMatch

Drug Development EDM Type

Description
EDM Sensitive type for detecting Covance Drug Development StudyID with ProposalID or ProtocolID.

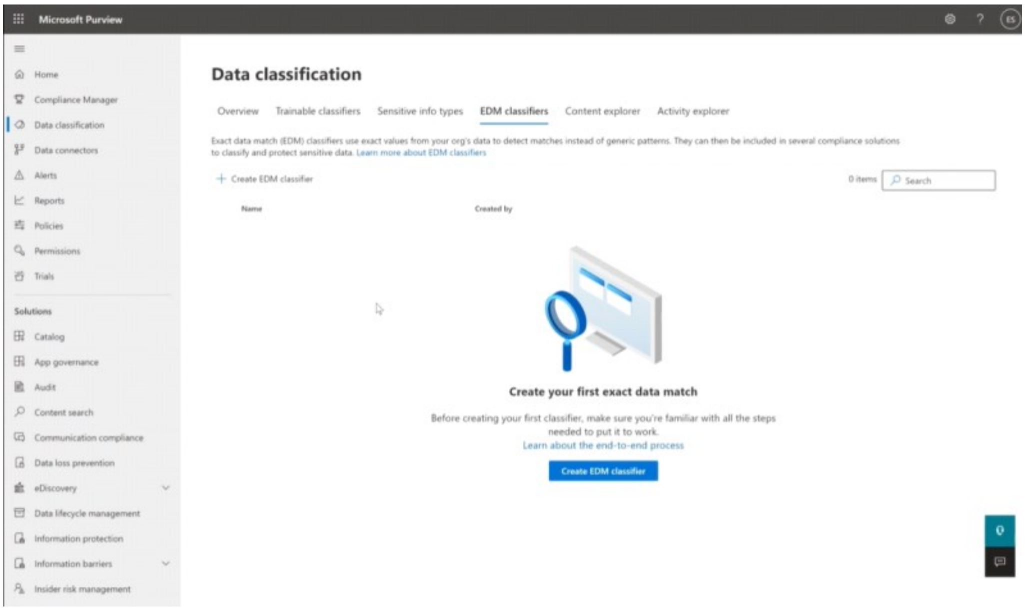
Confidence level
Medium

Created by
MIPDemo

Notes:

- This is in the regular SIT UI, not in the EDM UI!!!
- EDM sensitive info type changes take up to *one hour* to be propagated. You might be testing the old version!!!

Sneak peek: new EDM wizard

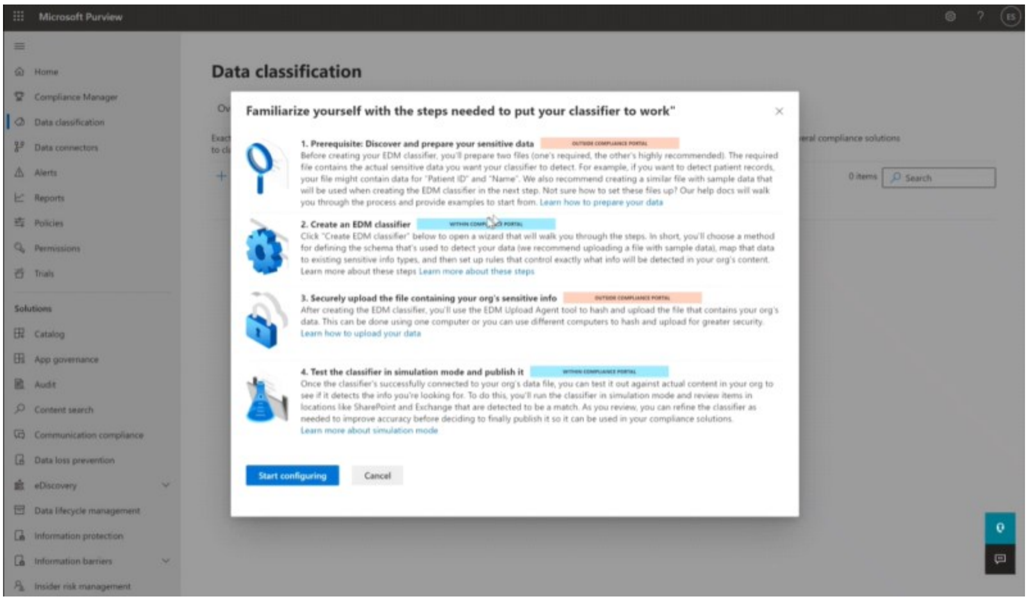


Data classification

Exact data match (EDM) classifiers use exact values from your org's data to detect matches instead of generic patterns. They can then be included in several compliance solutions to classify and protect sensitive data. Learn more about EDM classifiers

Create your first exact data match

Before creating your first classifier, make sure you're familiar with all the steps needed to put it to work.
Learn about the end-to-end process



Data classification

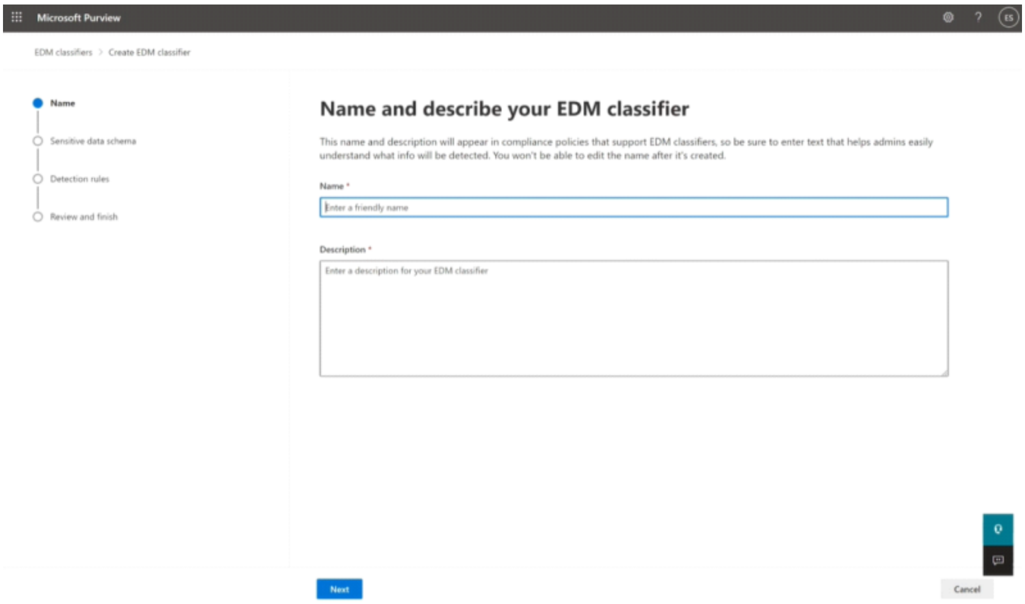
Familiarize yourself with the steps needed to put your classifier to work"

1. Prerequisite: Discover and prepare your sensitive data OUTSIDE COMPLIANCE PORTAL
Before creating your EDM classifier, you'll prepare two files (one's required, the other's highly recommended). The required file contains the actual sensitive data you want your classifier to detect. For example, if you want to detect patient records, your file might contain data for "Patient ID" and "Name". We also recommend creating a similar file with sample data that will be used when creating the EDM classifier in the next step. Not sure how to set these files up? Our help docs will walk you through the process and provide examples to start from. Learn how to prepare your data

2. Create an EDM classifier WITHIN COMPLIANCE PORTAL
Click "Create EDM classifier" below to open a wizard that will walk you through the steps. In short, you'll choose a method for defining the schema that's used to detect your data (we recommend uploading a file with sample data), map that data to existing sensitive info types, and then set up rules that control exactly what info will be detected in your org's content. Learn more about these steps Learn more about these steps

3. Securely upload the file containing your org's sensitive info OUTSIDE COMPLIANCE PORTAL
After creating the EDM classifier, you'll use the EDM Upload Agent tool to hash and upload the file that contains your org's data. This can be done using one computer or you can use different computers to hash and upload for greater security. Learn how to upload your data

4. Test the classifier in simulation mode and publish it WITHIN COMPLIANCE PORTAL
Once the classifier's successfully connected to your org's data file, you can test it out against actual content in your org to see if it detects the info you're looking for. To do this, you'll run the classifier in simulation mode and review items in locations like SharePoint and Exchange that are detected to be a match. As you review, you can refine the classifier as needed to improve accuracy before deciding to finally publish it so it can be used in your compliance solutions. Learn more about simulation mode

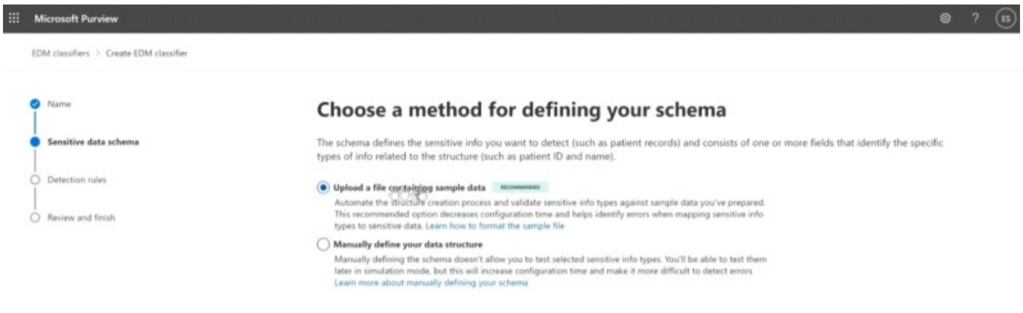


Name and describe your EDM classifier

This name and description will appear in compliance policies that support EDM classifiers, so be sure to enter text that helps admins easily understand what info will be detected. You won't be able to edit the name after it's created.

Name *
Enter a friendly name

Description *
Enter a description for your EDM classifier

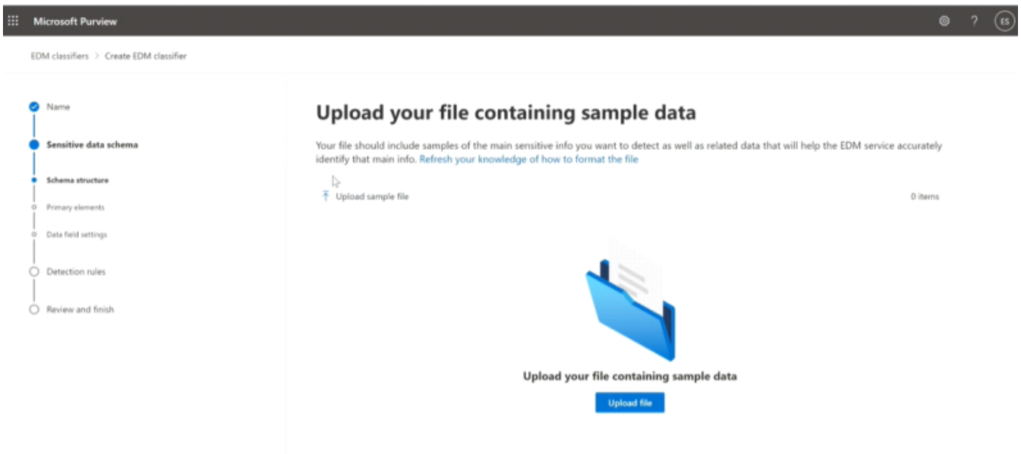


Choose a method for defining your schema

The schema defines the sensitive info you want to detect (such as patient records) and consists of one or more fields that identify the specific types of info related to the structure (such as patient ID and name).

● **Upload a file containing sample data** RECOMMENDED
Automate the structure creation process and validate sensitive info types against sample data you've prepared. This recommended option decreases configuration time and helps identify errors when mapping sensitive info types to sensitive data. Learn how to format the sample file

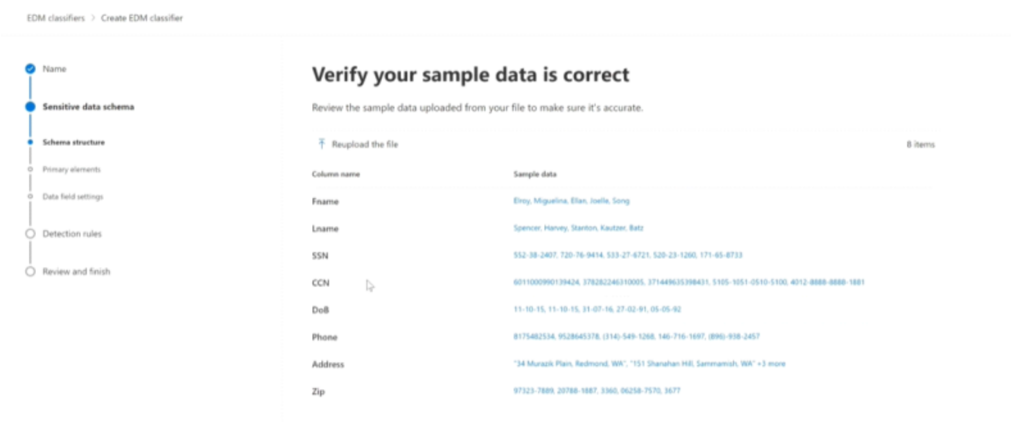
○ **Manually define your data structure**
Manually defining the schema doesn't allow you to test selected sensitive info types. You'll be able to test them later in simulation mode, but this will increase configuration time and make it more difficult to detect errors. Learn more about manually defining your schema



Upload your file containing sample data

Your file should include samples of the main sensitive info you want to detect as well as related data that will help the EDM service accurately identify that main info. Refresh your knowledge of how to format the file

Upload your file containing sample data

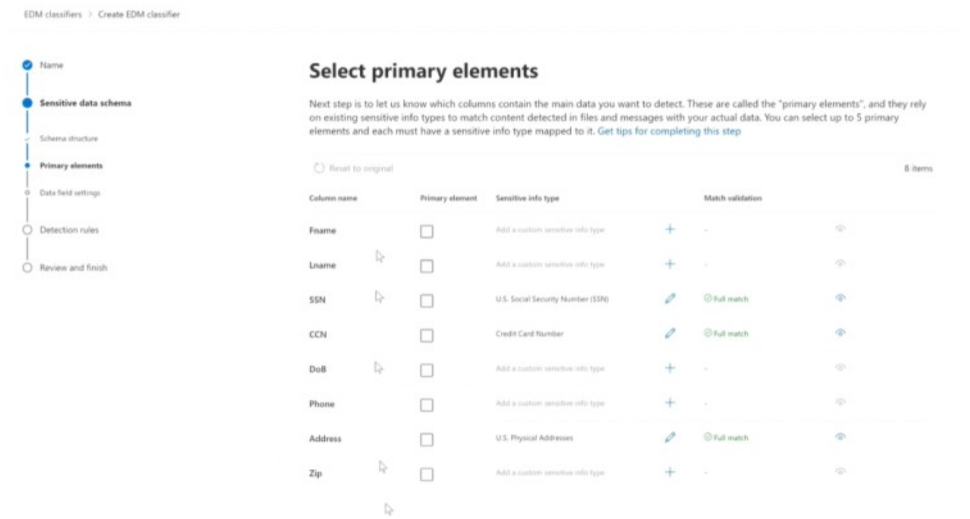


Verify your sample data is correct

Review the sample data uploaded from your file to make sure it's accurate.

8 items

Column name	Sample data
Fname	Elroy, Miguelina, Ellan, Joelle, Song
Lname	Spencer, Harvey, Stanton, Kautzer, Batz
SSN	552-38-2407, 720-76-9414, 533-27-6721, 520-23-1260, 171-65-8733
CCN	6011000990139424, 378282246310005, 371449635398431, 5105-1051-0510-5100, 4012-8888-8888-1881
DoB	11-10-15, 11-10-15, 31-07-16, 27-02-91, 05-05-92
Phone	8175482534, 9528645378, (314)-549-1268, 146-716-1697, (896)-938-2457
Address	"34 Murazik Plain, Redmond, WA", "151 Shanahan Hill, Sammamish, WA" +3 more
Zip	97323-7889, 20788-1887, 3360, 06258-7570, 3677

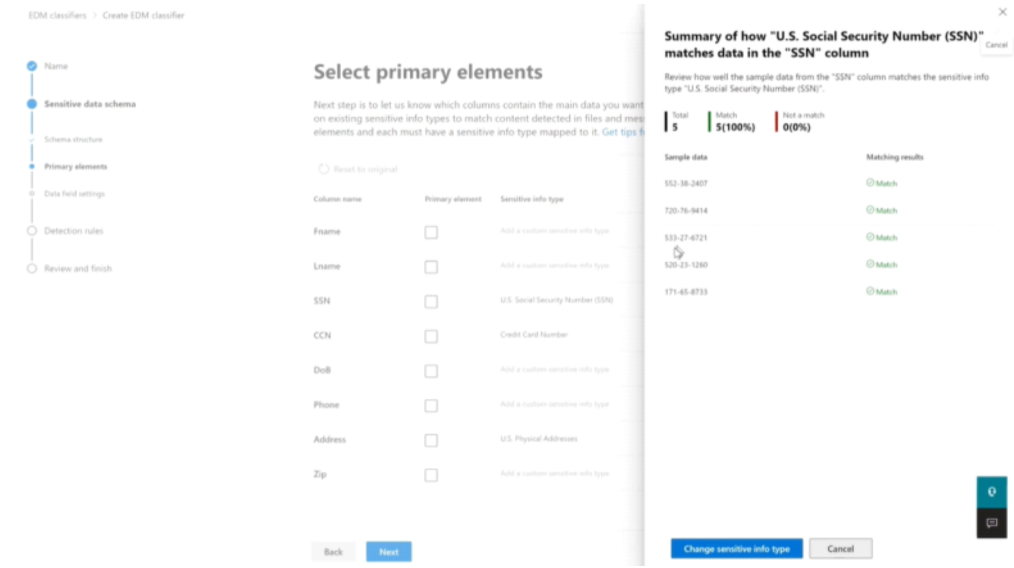


Select primary elements

Next step is to let us know which columns contain the main data you want to detect. These are called the "primary elements", and they rely on existing sensitive info types to match content detected in files and messages with your actual data. You can select up to 5 primary elements and each must have a sensitive info type mapped to it. Get tips for completing this step

8 items

Column name	Primary element	Sensitive info type		Match validation
Fname	□	Add a custom sensitive info type	+	-
Lname	□	Add a custom sensitive info type	+	-
SSN	□	U.S. Social Security Number (SSN)		✓ Full match
CCN	□	Credit Card Number		✓ Full match
DoB	□	Add a custom sensitive info type	+	-
Phone	□	Add a custom sensitive info type	+	-
Address	□	U.S. Physical Addresses		✓ Full match
Zip	□	Add a custom sensitive info type	+	-

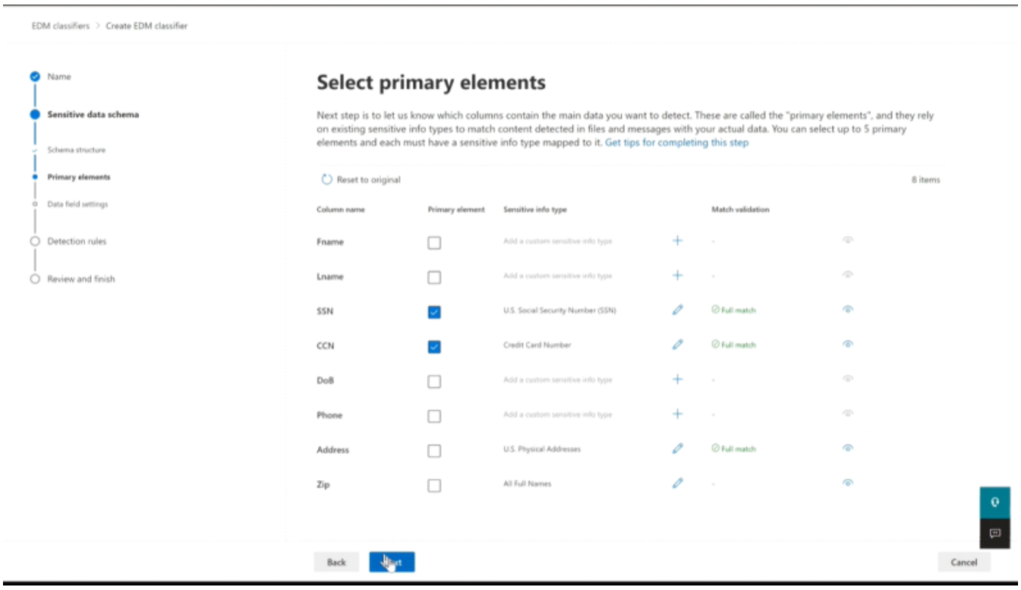


Summary of how "U.S. Social Security Number (SSN)" matches data in the "SSN" column

Review how well the sample data from the "SSN" column matches the sensitive info type "U.S. Social Security Number (SSN)".

Total	Match	Not a match
5	5(100%)	0(0%)

Sample data	Matching results
552-38-2407	✓ Match
720-76-9414	✓ Match
533-27-6721	✓ Match
520-23-1260	✓ Match
171-65-8733	✓ Match

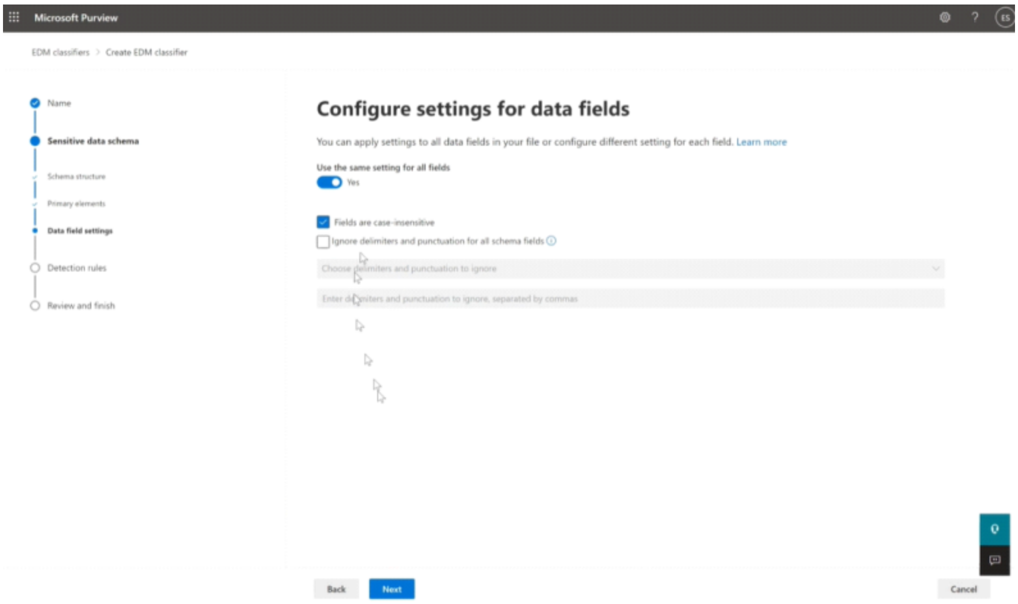


Select primary elements

Next step is to let us know which columns contain the main data you want to detect. These are called the "primary elements", and they rely on existing sensitive info types to match content detected in files and messages with your actual data. You can select up to 5 primary elements and each must have a sensitive info type mapped to it. Get tips for completing this step

Reset to original — 8 items

Column name	Primary element	Sensitive info type		Match validation
Fname	☐	Add a custom sensitive info type	+	-
Lname	☐	Add a custom sensitive info type	+	-
SSN	☑	U.S. Social Security Number (SSN)		✓ Full match
CCN	☑	Credit Card Number		✓ Full match
DoB	☐	Add a custom sensitive info type	+	-
Phone	☐	Add a custom sensitive info type	+	-
Address	☐	U.S. Physical Addresses		✓ Full match
Zip	☐	All Full Names		-



Configure settings for data fields

You can apply settings to all data fields in your file or configure different setting for each field. Learn more

Use the same setting for all fields

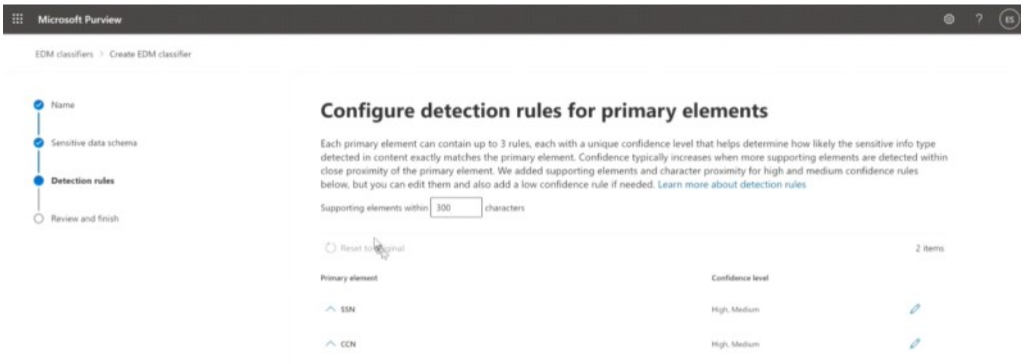
Yes

☑ Fields are case-insensitive

☐ Ignore delimiters and punctuation for all schema fields

Choose delimiters and punctuation to ignore

Enter delimiters and punctuation to ignore, separated by commas



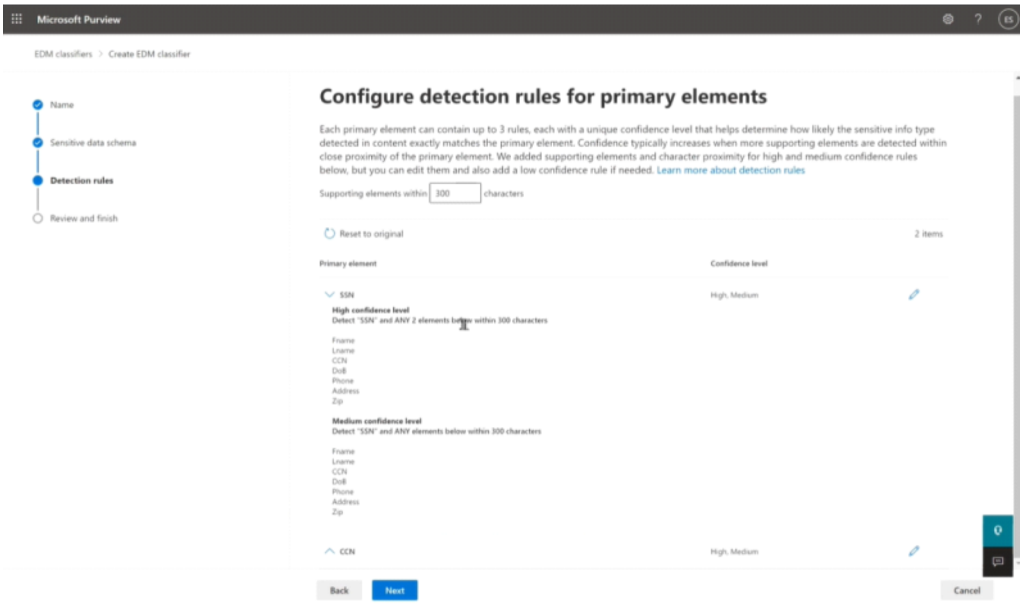
Configure detection rules for primary elements

Each primary element can contain up to 3 rules, each with a unique confidence level that helps determine how likely the sensitive info type detected in content exactly matches the primary element. Confidence typically increases when more supporting elements are detected within close proximity of the primary element. We added supporting elements and character proximity for high and medium confidence rules below, but you can edit them and also add a low confidence rule if needed. Learn more about detection rules

Supporting elements within 300 characters

Reset to original — 2 items

Primary element	Confidence level
SSN	High, Medium
CCN	High, Medium



Configure detection rules for primary elements

Each primary element can contain up to 3 rules, each with a unique confidence level that helps determine how likely the sensitive info type detected in content exactly matches the primary element. Confidence typically increases when more supporting elements are detected within close proximity of the primary element. We added supporting elements and character proximity for high and medium confidence rules below, but you can edit them and also add a low confidence rule if needed. Learn more about detection rules

Supporting elements within 300 characters

Reset to original — 2 items

Primary element	Confidence level
SSN	High, Medium
CCN	High, Medium

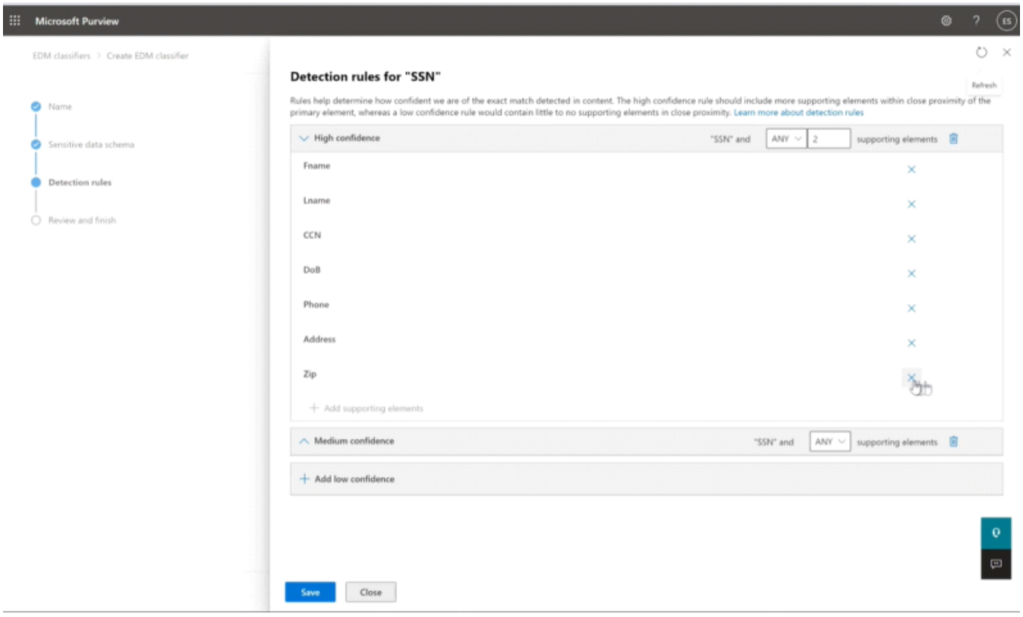
SSN:

High confidence level
Detect "SSN" and ANY 2 elements below within 300 characters

Fname
Lname
CCN
DoB
Phone
Address
Zip

Medium confidence level
Detect "SSN" and ANY elements below within 300 characters

Fname
Lname
CCN
DoB
Phone
Address
Zip



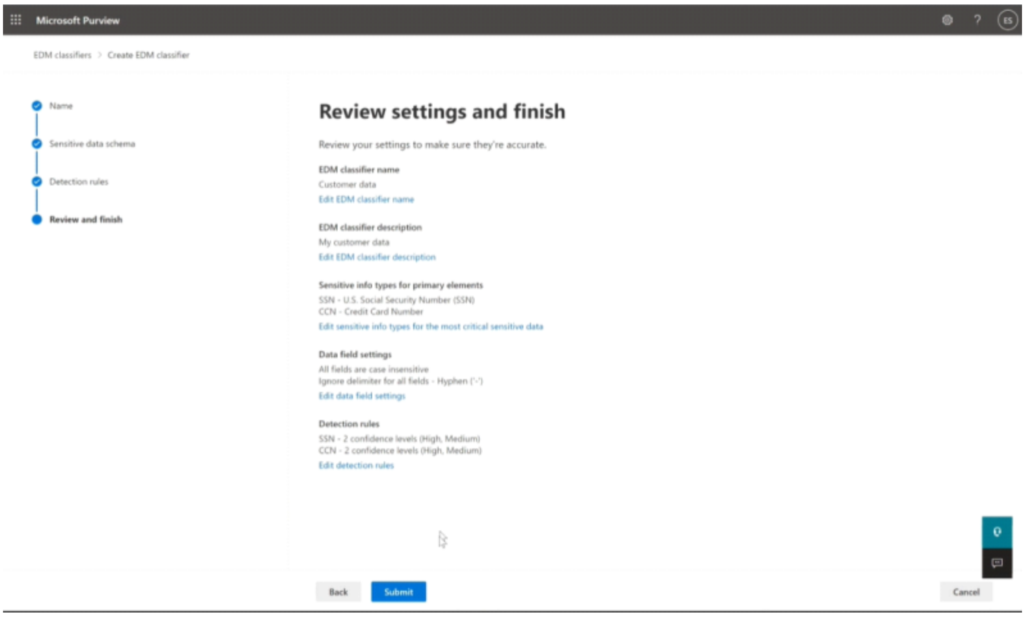
Detection rules for "SSN"

Rules help determine how confident we are of the exact match detected in content. The high confidence rule should include more supporting elements within close proximity of the primary element, whereas a low confidence rule would contain little to no supporting elements in close proximity. Learn more about detection rules

High confidence — "SSN" and ANY 2 supporting elements

Fname
Lname
CCN
DoB
Phone
Address
Zip

Medium confidence — "SSN" and ANY supporting elements



Review settings and finish

Review your settings to make sure they're accurate.

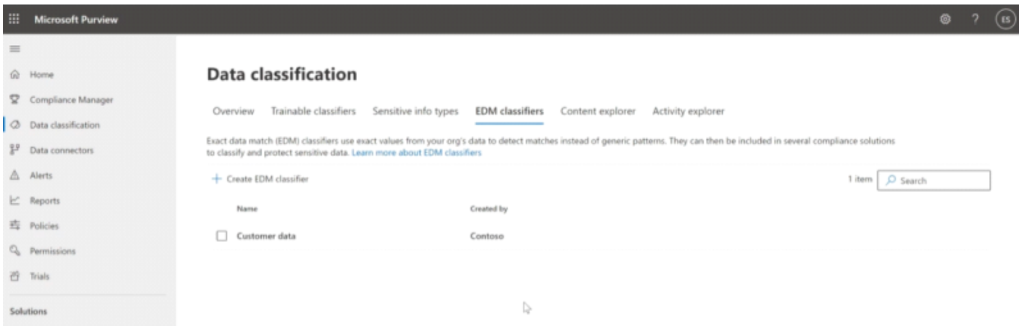
EDM classifier name
Customer data
Edit EDM classifier name

EDM classifier description
My customer data
Edit EDM classifier description

Sensitive info types for primary elements
SSN - U.S. Social Security Number (SSN)
CCN - Credit Card Number
Edit sensitive info types for the most critical sensitive data

Data field settings
All fields are case insensitive
Ignore delimiter for all fields - Hyphen ('-')
Edit data field settings

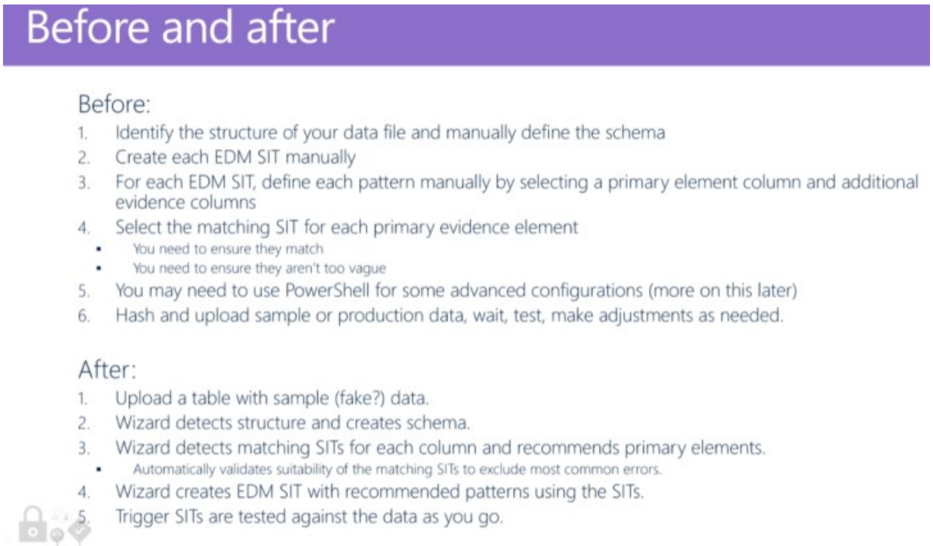
Detection rules
SSN - 2 confidence levels (High, Medium)
CCN - 2 confidence levels (High, Medium)
Edit detection rules



Data classification

Exact data match (EDM) classifiers use exact values from your org's data to detect matches instead of generic patterns. They can then be included in several compliance solutions to classify and protect sensitive data. Learn more about EDM classifiers

	Name	Created by
□	Customer data	Contoso



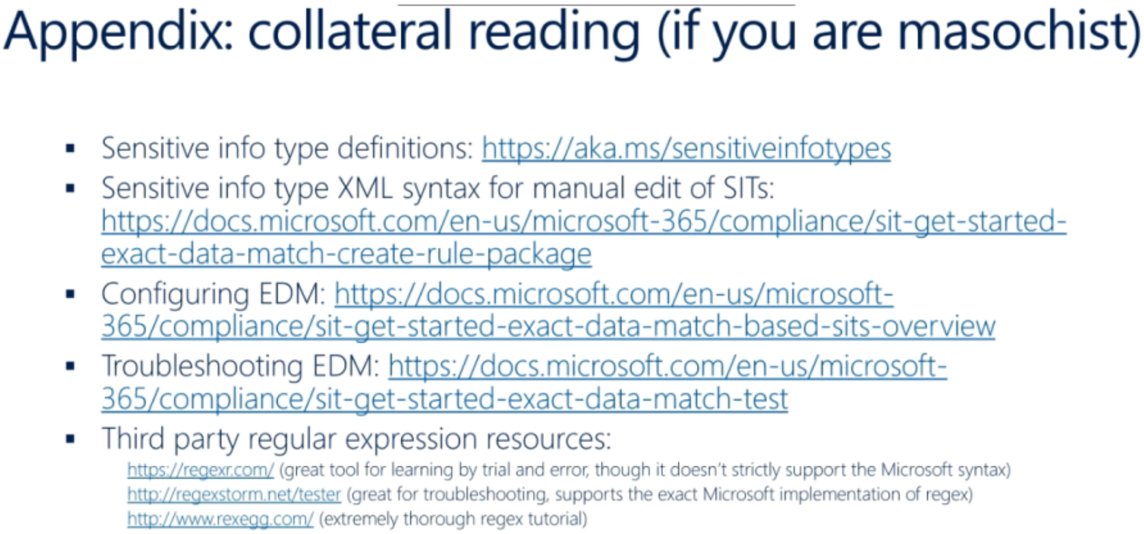
Before and after

Before:

1. Identify the structure of your data file and manually define the schema
2. Create each EDM SIT manually
3. For each EDM SIT, define each pattern manually by selecting a primary element column and additional evidence columns
4. Select the matching SIT for each primary evidence element
   - You need to ensure they match
   - You need to ensure they aren't too vague
5. You may need to use PowerShell for some advanced configurations (more on this later)
6. Hash and upload sample or production data, wait, test, make adjustments as needed.

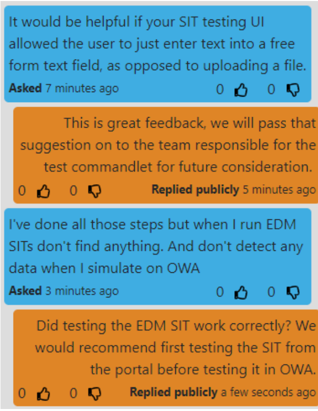
After:

1. Upload a table with sample (fake?) data.
2. Wizard detects structure and creates schema.
3. Wizard detects matching SITs for each column and recommends primary elements.
   - Automatically validates suitability of the matching SITs to exclude most common errors.
4. Wizard creates EDM SIT with recommended patterns using the SITs.
5. Trigger SITs are tested against the data as you go.



Appendix: collateral reading (if you are masochist)

- Sensitive info type definitions: https://aka.ms/sensitiveinfotypes
- Sensitive info type XML syntax for manual edit of SITs: https://docs.microsoft.com/en-us/microsoft-365/compliance/sit-get-started-exact-data-match-create-rule-package
- Configuring EDM: https://docs.microsoft.com/en-us/microsoft-365/compliance/sit-get-started-exact-data-match-based-sits-overview
- Troubleshooting EDM: https://docs.microsoft.com/en-us/microsoft-365/compliance/sit-get-started-exact-data-match-test
- Third party regular expression resources:
  https://regexr.com/ (great tool for learning by trial and error, though it doesn't strictly support the Microsoft syntax)
  http://regexstorm.net/tester (great for troubleshooting, supports the exact Microsoft implementation of regex)
  http://www.rexegg.com/ (extremely thorough regex tutorial)



It would be helpful if your SIT testing UI allowed the user to just enter text into a free form text field, as opposed to uploading a file.
Asked 7 minutes ago 0 👍 0 👎

This is great feedback, we will pass that suggestion on to the team responsible for the test commandlet for future consideration.
0 👍 0 👎 **Replied publicly** 5 minutes ago

I've done all those steps but when I run EDM SITs don't find anything. And don't detect any data when I simulate on OWA
Asked 3 minutes ago 0 👍 0 👎

Did testing the EDM SIT work correctly? We would recommend first testing the SIT from the portal before testing it in OWA.
0 👍 0 👎 **Replied publicly** a few seconds ago