Broadcom CA Test Data Manager HIPAA Classifier

Problem Statement:

CA Test Data Manager provides the ability to scan data sources for privacy information. A number of out-of-the-box classifiers are delivered with the product. A given organization may have requirements to find data beyond the standard set (e.g. an Account Number configured in a certain way), or finds the OOTB set of classifiers too broad leading to large numbers of "false posititives". This document describes how the classifier framework can be extended for custom purposes by using a real-life example.

HIPAA

In the US, the Health Insurance Portability and Accountability Act of 1996 ("HIPAA") includes a Privacy Rule providing guidance on **Protected Health Information (PHI)**. https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

(1) the geographic units formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and

(2) the initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000;

(C) All elements of dates (except year) for dates directly related to the individual, including birth date, admission date, discharge date, date of death;

and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;

(D) Telephone numbers;

(E) Fax numbers;

(F) Electronic mail addresses:

(G) Social security numbers;

(H) Medical record numbers;

(I) Health plan beneficiary numbers;

(J) Account numbers;

(K) Certificate/license numbers;

(L) Vehicle identifiers and serial numbers, including license plate numbers;

(M) Device identifiers and serial numbers;

(N) Web Universal Resource Locators (URLs);

(O) Internet Protocol (IP) address numbers;

(P) Biometric identifiers, including finger and voice prints;

(Q) Full face photographic images and any comparable images; and any other unique identifying number, characteristic, or code, except as permitted for re-identification purposes provided certain conditions are met.

The HIPAA classifier described here helps organizations with the identification of candidate fields that may contain data associated with the 18 items listed. Once this information is audited/verified, CA TDM's Masking utilities may be used to mask/obfuscate/sanitize the information so that production copies of data may be used in lower environments while protecting the Privacy of the original PHI information.

The requirements for scanning have been collected from a variety of sources. These inputs, and the tests run to validate them are documented below for audit purposes.

(1) **the geographic units formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and**

(2) the initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000;

This component of the HIPAA classifier is a Seedlist classifier. It scans for specific Zip Codes using a seedlist of specific Zip Codes (first 3 digits) that contain 20,000 or fewer people. The Seedlist should be registered with the name "Zip Codes under 20000 population".

Rather than re-do the calculations from the publicly available information from the US Census website, this list was retrieved from https://www.johndcook.com/blog/2016/06/29/sparsely-populated-zip-codes/

ZipCode
036
059
063
102
203
556
692
790
821
823
830
831
878
879
884
890
893

(C) All elements of dates (except year) for dates directly related to the individual, including birth date, admission date, discharge date, date of death;

and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;

This component of the HIPAA classifier is a Content classifier. It scans for Dates using Regular Expressions (Regex) strings. The strings include:

"((?:19|20)\d\d)([- /.])(0?[1-9]|1[012])\2(0?[1-9]|[12]\d|3[01])"

This regex string matches Dates with the following formats:

2004/01/01

2004-01-01

1994/01/01

1994-01-01

"((?:19|20)\d\d)(0[1-9]|1[012])(0[1-9]|[12]\d|3[01])"

This regex string matches Dates with the following formats:

19000101

20040229

"(0?[1-9]|1[012])([- /.])(0?[1-9]|[12]\d|3[01])\2((?:19|20)\d\d)"

This regex string matches Dates with the following formats:

01/01/2004

01-01-2004

01/01/1994

01-01-1994

"^\d\d?-\w\w\w-\d\d\d\d"

This regex string matches Dates with the following formats:

01-JAN-2018

It does not check for a valid date, just the pattern.

(D) Telephone numbers;

This component of the HIPAA classifier is a Content classifier. It scans for US telephone numbers using Regular Expressions (Regex) strings. The strings include:

"^1?\s*\W?\s*([2-9][0-8][0-9])\s*\W?\s*([2-9][0-9]{2})\s*\W?\s*([0-9]{4})(\se?x?t?(\d*))?"

This regex string matches Telephone Numbers with the following formats:

1 456 789 1010

1-456-789-1010

1 (456) 789-1010

(456) 789-1010

(456) 789 1010

456-789-1010

456.789.1010

456 789 1010

456 789 1010 x123

(E) Fax numbers;

This component of the HIPAA classifier is a Content classifier. It scans for (primarily US) fax numbers using Regular Expressions (Regex) strings. The strings include:

"^1?\s*\W?\s*([2-9][0-8][0-9])\s*\W?\s*([2-9][0-9]{2})\s*\W?\s*([0-9]{4})(\se?x?t?(\d*))?"

This regex string matches Telephone Numbers with the following formats:

1 456 789 1010

1-456-789-1010

1 (456) 789-1010

(456) 789-1010

(456) 789 1010

456-789-1010

456.789.1010

456 789 1010

456 789 1010 x123

(F) Electronic mail addresses:

This component of the HIPAA classifier is a Content classifier. It scans for email addresses using a Regular Expression (Regex) string. The strings include:

"^[a-zA-Z0-9.!#$%&'*+\/=?^_`{|}~-]+@[a-zA-Z0-9](?:[a-zA-Z0-9-]{0,61}[a-zA-Z0-9])?(?:\.[a-zA-Z0-9](?:[a-zA-Z0-9-]{0,61}[a-zA-Z0-9])?)*$"

This regex string matches Electronic Mail Addresses with the following formats:

email@example.com

firstname.lastname@example.com

email@subdomain.example.com

firstname+lastname@example.com

email@123.123.123.123

1234567890@example.com

email@example-one.com

_____@example.com

email@example.name

email@example.museum

email@example.co.jp

firstname-lastname@example.com

(G) Social security numbers;

This component of the HIPAA classifier is a Content classifier. It scans for Social Security Numbers using Regular Expressions (Regex) strings. The strings include:

"(?!(?:666)|(?:000))(?:(?!9)(?:\d)(?:\d{2}))(?:-?(?!(?:00))(?:\d{2})){3}"

This regex string matches valid SSNs with the following formats:

123-45-6789

123456789

(H) Medical record numbers;

This component of the HIPAA classifier is a Content classifier. It scans for Medical Record Numbers using Regular Expressions (Regex) strings. The strings include:

"\d{10}" **This will depend on the organization's definition of a Medical Record Number. There is no industry standardization of this information. Customize this regex string for the particular usage.**

This regex string matches Medical Record Numbers with the following formats:

1234567890

(I) Health plan beneficiary numbers;

This component of the HIPAA classifier is a Content classifier. It scans for Health Plan Beneficiary Numbers using Regular Expressions (Regex) strings. The strings include:

"\d{10}" **This will depend on the organization's definition of a Health Plan Beneficiary Number. There is no industry standardization of this information. Customize this regex string for the particular usage.**

This regex string matches Health Plan Beneficiary Numbers with the following formats:

1234567890

(J) Account numbers;

This component of the HIPAA classifier is a Content classifier. It scans for Account Numbers using Regular Expressions (Regex) strings. The strings include:

"\d{10}" **This will depend on the organization's definition of an Account Number. There is no industry standardization of this information. Customize this regex string for the particular usage.**

This regex string matches Account Numbers with the following formats:

1234567890

(K) Certificate/license numbers;

This component of the HIPAA classifier is a Content classifier. It scans for Certificate/License Numbers using Regular Expressions (Regex) strings. The strings include:

"\d{10}" **This will depend on the organization's definition of a Certificate/License Number. There is no industry standardization of this information. Customize this regex string for the particular usage.**

This regex string matches Certificate/License Numbers with the following formats:

1234567890

(L) Vehicle identifiers and serial numbers, including license plate numbers;

This component of the HIPAA classifier is a Content classifier. It scans for 17-digit VINs and 7-character License Plates using Regular Expressions (Regex) strings. The strings include:

17-digit VIN:

"^([A-HJ-NPR-Z0-9]{17})"

7-character License Plate:

"^([A-Z0-9]{7})"

(M) Device identifiers and serial numbers;

This component of the HIPAA classifier is a Content classifier. It scans for device UDIs and Serial Numbers using Regular Expressions (Regex) strings. The strings include:

"\d{14}" (Using 14-digit human readable DI examples from webcast on FDA site)

UDI:

Device Identifier followed by the Production Identifier

- a device identifier (DI), a mandatory, fixed portion of a UDI that identifies the labeler and the specific version or model of a device, and
- a production identifier (PI), a conditional, variable portion of a UDI that identifies one or more of the following when included on the label of a device:
  - the lot or batch number within which a device was manufactured;
  - the serial number of a specific device;
  - the expiration date of a specific device;
  - the date a specific device was manufactured;
  - the distinct identification code required by §1271.290(c) for a human cell, tissue, or cellular and tissue-based product (HCT/P) regulated as a device.

(N) Web Universal Resource Locators (URLs);

This component of the HIPAA classifier is a Content classifier. It scans for URLs using a Regular Expressions (Regex) string. The string includes:

"^(?:http(s)?:\/\/)?[\w.-]+(?:\.[\w\.-]+)+[\w\-\._~:/?#[\]@!\$&'\(\)*\+,;=.]+$"

This regex string matches URLs with the following formats:

www.ca.com

ca.com

http://www.ca.com

https://www.ca.com

docops.ca.com

localhost

10.10.20.20

ca.com/tdm

(O) Internet Protocol (IP) address numbers;

This component of the HIPAA classifier is a Content classifier. It scans for IP addresses using a Regular Expressions (Regex) string. The string includes:

IPv4:

"((25[0-5])|(2[0-4][0-9])|([01]?[0-9][0-9]?))\.((25[0-5])|(2[0-4][0-9])|([01]?[0-9][0-9]?))\.((25[0-5])|(2[0-4][0-9])|([01]?[0-9][0-9]?))\.((25[0-5])|(2[0-4][0-9])|([01]?[0-9][0-9]?))"

This regex string matches IPv4 Addresses with the following formats:

192.168.1.1

10.10.10.10

IPv6:

"((?:(?:[A-F])|(?:\d)){1,4}::?){4,7}(?:(?:(?:[A-F])|(?:\d)){1,4})(?:(?:(?:%)|(?:/))?(?:[A-F\d]{0,4}))"

This regex string matches IPv6 Addresses with the following formats:

2001:db8:3:4::192.168.1.1

fe80::95e8:a899:dca9:3cf6%17

(P) Biometric identifiers, including finger and voice prints;

As this information is usually in a Binary or BLOB format, a Content Classifier does not apply. If specific column names are known, then a Column Classifier may apply. As an example:

First RegEx is

".*(?:Fi?nger?.*Pr?i?nt|Pr?in?t).*”

Matches the following column names:

Finger_Print

FingerPrint

Print

Prints

Second RegEx is

“.*(?:Vo?ice?.*Pr?i?nt|Pr?in?t).*”

Matches the following column names:

Voice_Print

VoicePrint

Print

Prints

(Q) Full face photographic images and any comparable images; and any other unique identifying number, characteristic, or code, except as permitted for re-identification purposes provided certain conditions are met.

As image information is usually in a Binary or BLOB format, a Content Classifier does not apply. If specific column names are known, then a Column Classifier may apply. As an example:

".*(?:Ph?o?to?.*gr?a?ph|Ph?ot?o).*"

Matches the following column names:

Photo

Photograph

If this applies to an "Employee Number", then use a Content Classifier.

Creation of the Classifier Package

Now that the definition of the seedlists or regular expressions has been completed, we need to create the physical structure of the Classifier Package that we'll be importing into CA Test Data Manager.

Start by creating a folder with the Name being how you wish to identify the Classifier. We'll call ours HIPAA. Within that folder, create a separate .json file for each of the Classifiers. As an example: IP+Addresses.json


```
{
   "name":"IP Addresses",
   "description":"Identifies candidate data containing Personal Health Information",
   "classifierOrigin":"CA Technologies PreSales",
   "classifierClass":"com.ca.tdm.profiler.classifiers.RegExClassifier",
   "classifierType":"content",
   "tags":"IP Addresses",
   "config":[
                        {
      "name":"IPv4 Address",
      "value":"((25[0-5])|(2[0-4][0-9])|([01]?[0-9][0-9]?))\\.((25[0-5])|(2[0-4][0-9])|([01]?[0-9][0-9]?))\\.((25[0-5])|(2[0-4][0-9])|([01]?[0-9][0-9]?))\\.((25[0-5])|(2[0-4][0-9])|([01]?[0-9][0-9]?))"
                        },
                        {
      "name":"IPv6 Address",
      "value":"((?:(?:[A-F])|(?:\\d)){1,4}::?){4,7}(?:(?:(?:[A-F])|(?:\\d)){1,4})(?:(?:(?:%)|(?:/))?(?:[A-F\\d]{0,4}))"
      }
   ]
}
```

Classifiers may also contain a relationship from the Tag to the desired Masking algorithm to obfuscate this PII/PHI information. If your desired use case utilizes the TDM Portal end-to-end Masking scenario, then inclusion of the masking information into the classifier expedites the process. As an example: Account+Numbers.json


```
{
   "name":"Account Numbers",
   "description":"Identifies candidate data containing Personal Health Information",
   "classifierOrigin":"CA Technologies PreSales",
   "classifierClass":"com.ca.tdm.profiler.classifiers.RegExClassifier",
   "classifierType":"content",
   "tags":"Account Numbers",
   "config":[
                              {
          "name":"Account_Numbers",
          "value":"[\\d]{10}"
        }
      ],
  "maskFunctionGroup":

          [
                    {

                    "groupName": "masking functions",

                    "maskFunction":[
              {
                "functionName": "FORMATENCRYPT",
                "displayName": "FORMATENCRYPT",
                "notes": "Consistently masks the given column values by preserving the original format and keeping the unique values.",
```

```
            "maskParams": [
                    {
                      "paramPosition": "1",
                      "paramValue": "0"
                    },
                    {
                      "paramPosition": "2",
                      "paramValue": "0"
                    },
                                                                    {
                      "paramPosition": "3",
                      "paramValue": "10"
                    },
                                                                    {
                      "paramPosition": "4",
                      "paramValue": "0"
                    }
                  ]
              }
            ]
                  }
          ]
}
```

Notes and Cautions:

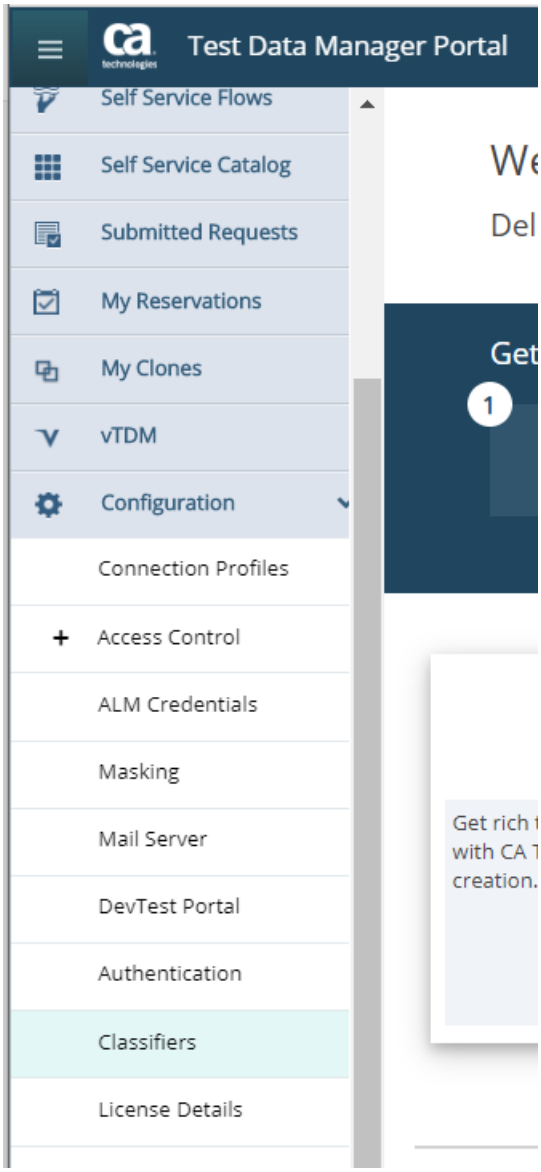
You may note that each of the .json file names has the names with a “+” sign in them. This is a way to have a descriptive name but also make the files importable.

The regex expressions need to be modified / tweaked to work within the Classifier framework. From the Account Number description, we stated that the regex string to match a 10-digit numeric was “\d{10}”. You’ll note in the above classifier .json, the string is now: "[\\d]{10}"

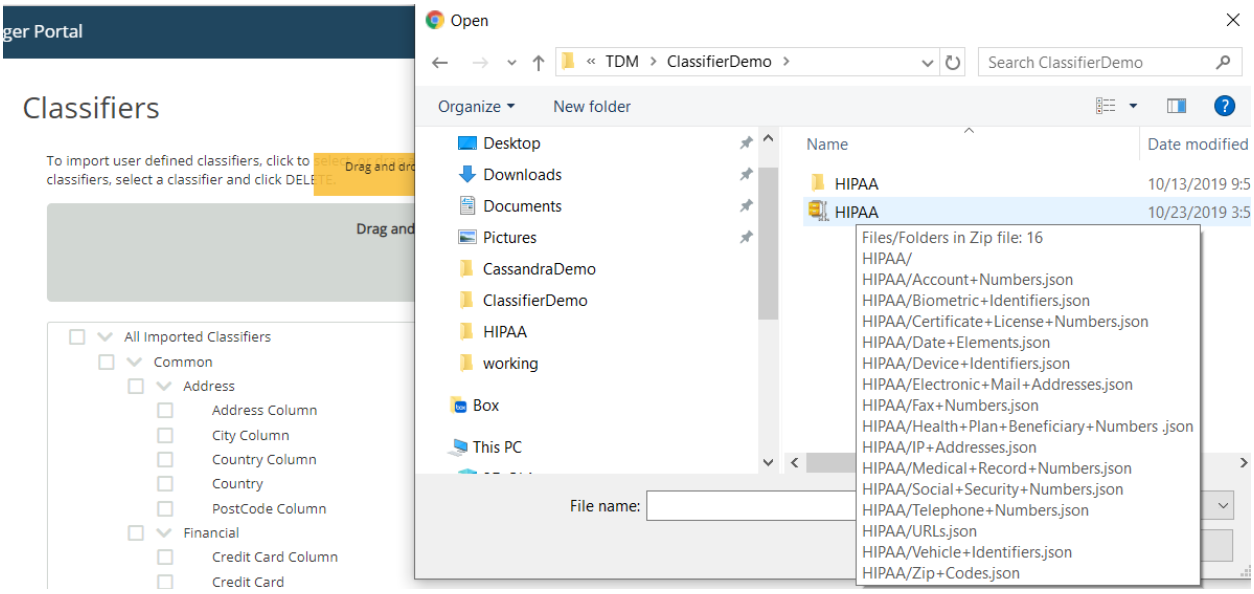
Once you've created all the .json files in the directory, go up one level and zip up the Directory.

HIPAA	10/13/2019 9:54 AM	File folder
HIPAA	10/23/2019 3:58 PM	WinZip File

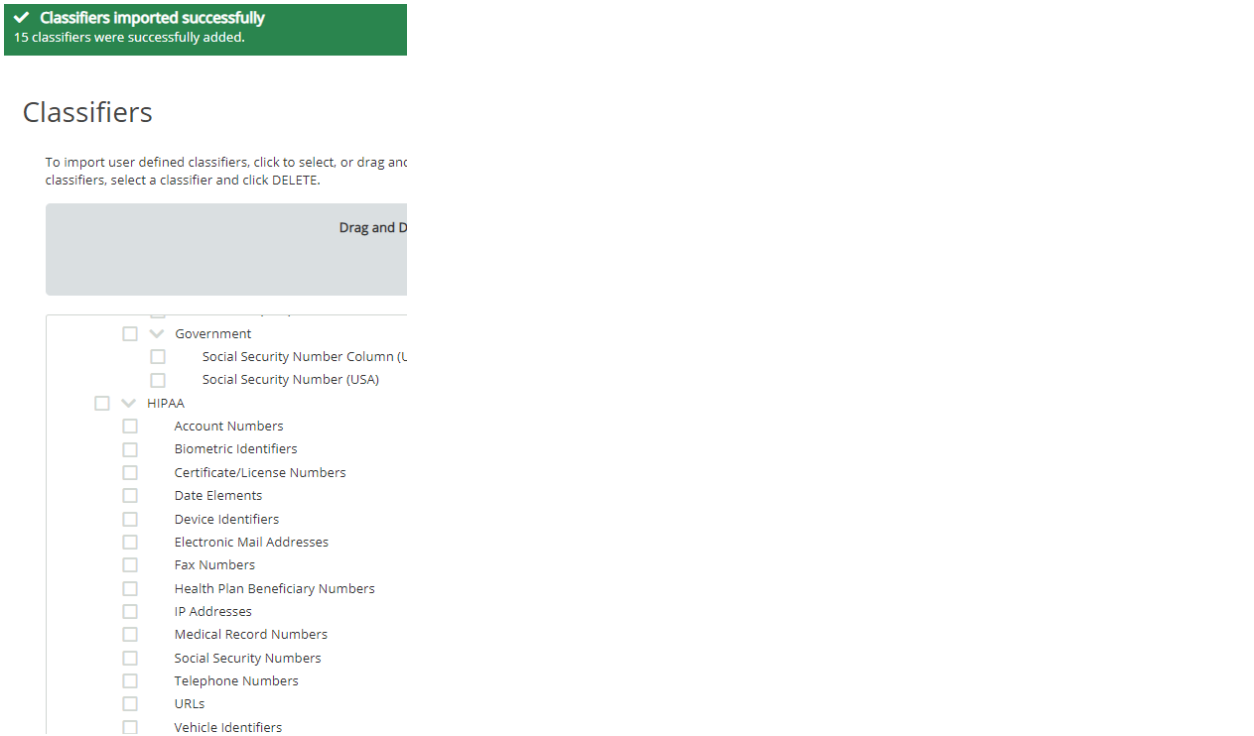
Proceed to the TDM Portal instance, and navigate to the Classifiers subfolder under Configuration



Drag-n-Drop the Classifier zip file (or select by browsing) into the import area.

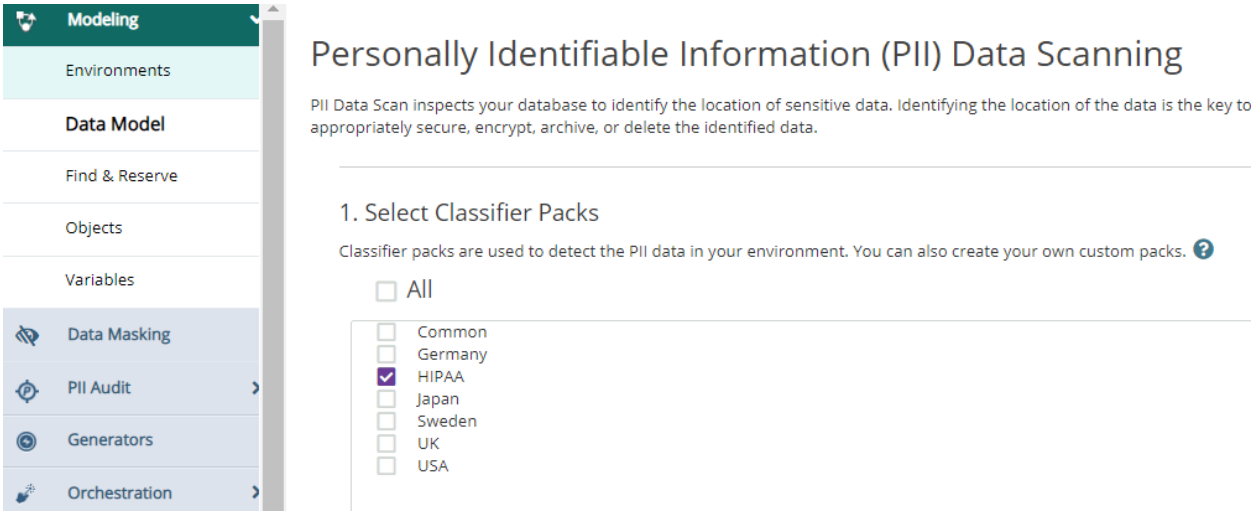


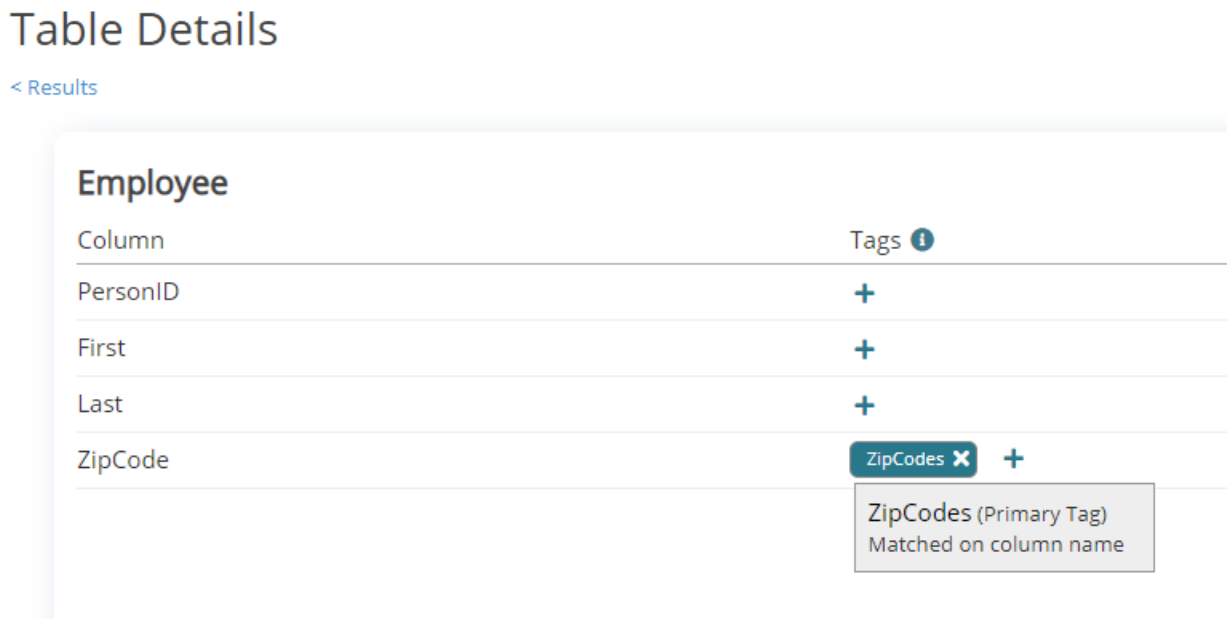
You will see an import dialog, followed by a Success message. Your classifiers will be imported



If any errors occur, review your file structures again. Check the TDMModelService.log for any messages.

Now, when you select PII Scan from the Data Model dialog, the HIPAA classifier will appear





Disclaimers

This package and example are provided as-is. It is not complete without modification for a particular organization's data characteristics. This package is not supported by Broadcom.