Tot plugin Azure Storage

Control de versiones

Versión	Fecha de modificación	Responsable	Aprobador	Resumen de cambios
1.0	28/10/2022	Anjana Producto	Anjana Producto	Creación del documento

Modelo de Integración

Extracción de metadatos

Mediante las librerías ofrecidas por Azure, se autentica en la cuenta de almacenamiento que contiene el contenedor que se quiere gobernar.

Una vez realizada la conexión se recorre el contenedor en cuestión para poder generar un árbol representando todo el contenido del mismo.

Para la extracción de metadata de un objeto en cuestión se utiliza la misma conexión y las mismas herramientas proporcionadas por la librería de Azure para poder leer el metadata que posteriormente será enviado a Anjana para crear el objeto.

Extrae los siguientes atributos que deben llamarse igual en la tabla attribute_definition, campo name para que aparezcan en la plantilla.

- **schema** con el valor del contenedor de Azure.
- **physicalName** y name con el mismo valor, el nombre del fichero dentro de Azure.
- **path** con el path y el nombre del recurso si es un fichero.
- **infrastructure** con el valor seleccionado
- **technology** con el valor seleccionado
- **zone** con el valor seleccionado

También nos enviará los siguientes atributos relativos a los campos del recurso pedido, siempre dependiendo del contenido al que se refiere el recurso.

- CSV (.csv)
    - **name** con el valor del campo correspondiente
    - **fieldDataType** con el tipo de dato definido para el campo correspondiente
    - **position** posición que ocupa el campo correspondiente
- AVRO (.avro)
    - **name** con el valor del campo correspondiente
    - **defaultValue** con el valor por defecto definido para el campo correspondiente (si procede)
    - **fieldDataType** con el tipo de dato definido para el campo correspondiente
    - **position** posición que ocupa el campo correspondiente
    - **description** con el valor correspondiente para el campo (si procede)
- EXCEL (.xls .xlsx)
    - **name** con el valor del campo correspondiente

- **fieldDataType** con el tipo de dato definido para el campo correspondiente
- **position** posición que ocupa el campo correspondiente

- PARQUET (.parquet)
  - **name** con el valor del campo correspondiente
  - **fieldDataType** con el tipo de dato definido para el campo correspondiente
  - **position** posición que ocupa el campo correspondiente
  - **nullable** indicando si el campo correspondiente es nullable
  - **optional** indicando si el campo es opcional o no (si permite repetición)

Muestreo de datos

Mediante las librerías ofrecidas por Azure, se autentica en la cuenta de almacenamiento que contiene el contenedor que se quiere gobernar.

Una vez realizada la conexión, se localiza el objeto a muestrear, se lee su contenido (hasta el máximo número de resultados configurado) utilizando librerías de Apache según tipo de fichero qué es y se devuelven los resultados.

Gobierno activo

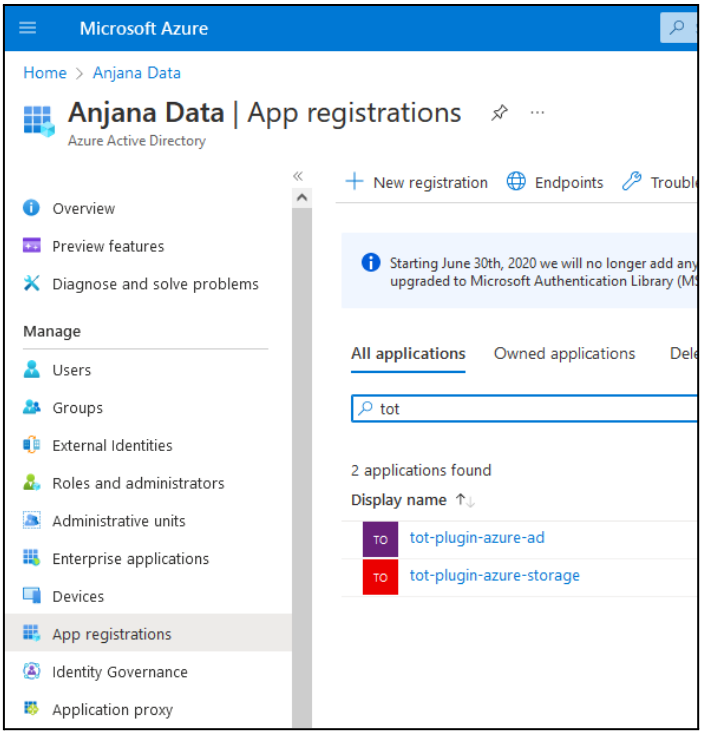
Mediante las librerías ofrecidas por Azure, se autentica en la cuenta de almacenamiento que contiene el contenedor que se quiere gobernar.

Se localiza el objeto en el que se quieren manipular los permisos, y mediante las herramientas de manipulación de ACLs, se añaden los permisos necesarios al grupo sobre los objetos (Read, Execute en todos los niveles desde el raíz hasta el fichero que representa el dataset y en el caso de representar un fichero particionado, a todos los ficheros presentes en ese momento) cuando se incluyen en un DSA. Adicionalmente, eliminar la ACL de un grupo cuando éste expira o el propio objeto expira.

Credenciales requeridas

Es necesario registrar una aplicación en Azure AD y generar el necesario clientID y secret para que el plugin pueda autenticar y adquirir los permisos necesarios para cada funcionalidad.

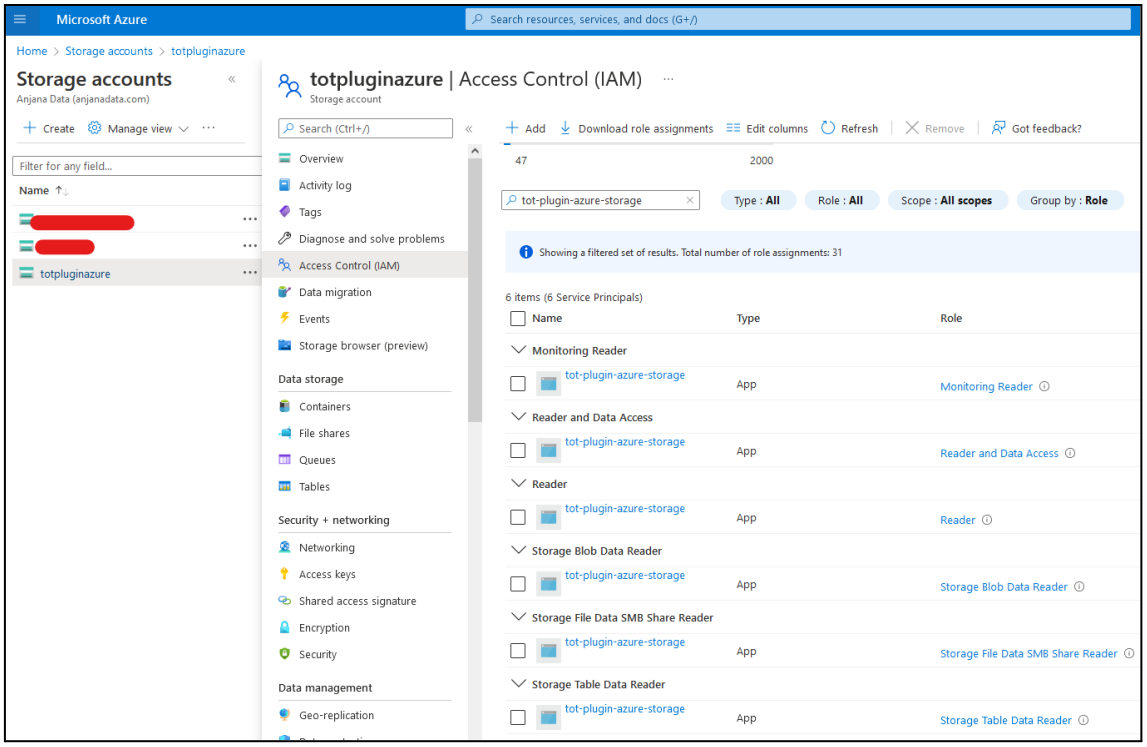
El plugin tiene la capacidad de manejar solo un storage account, por tanto, es necesario hacer una instancia para cada uno de los storage accounts a manejar.



Extracción de metadatos y muestreo de datos

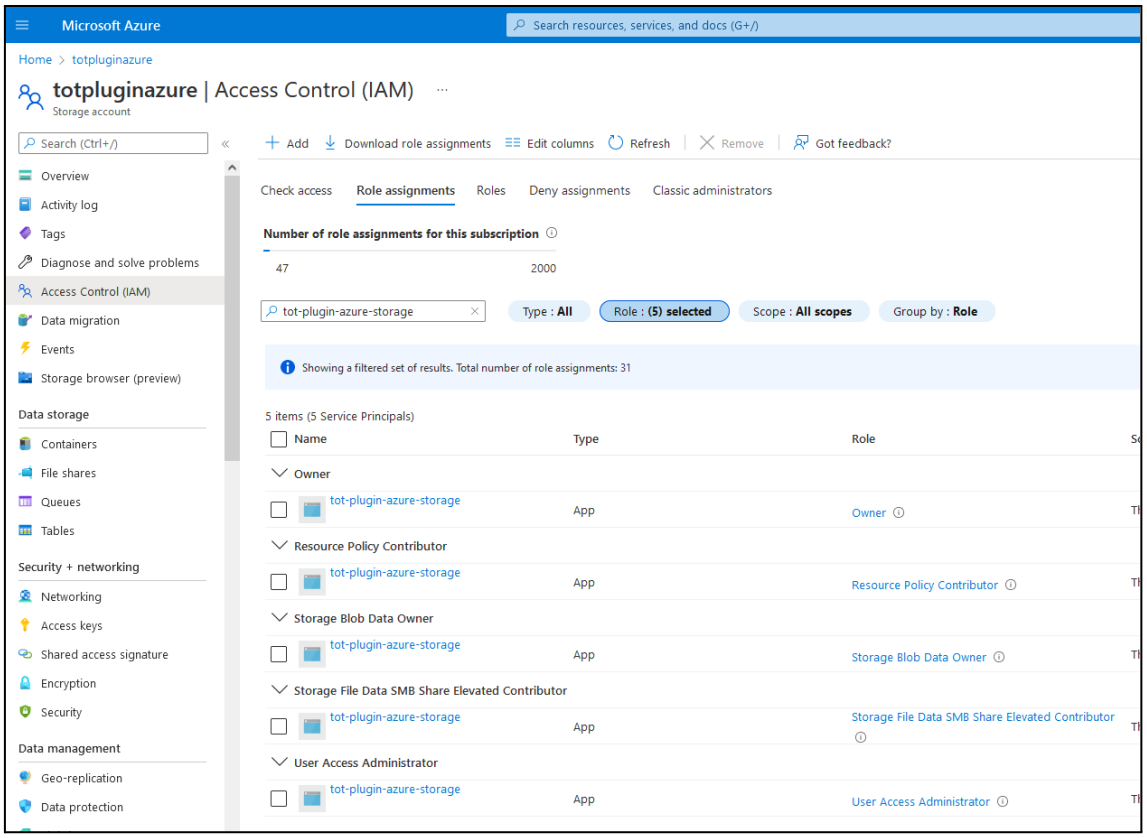
Se requiere permisos de lectura sobre las diferentes tipologías de almacenamiento más la propia configuración general de la cuenta de almacenamiento.

- Reader: View all resources, but does not allow you to make any changes.
- Storage Blob Data Reader: Allows for read access to Azure Storage blob containers and data
- Storage File Data SMB Share Reader: Allows for read access to Azure File Share over SMB
- Storage Table Data Reader: Allows for read access to Azure Storage tables and entities



Gobierno activo de accesos y estructuras

- User Access Administrator: Lets you manage user access to Azure resources.
- Storage Blob Data Owner: Allows for full access to Azure Storage blob containers and data, including assigning POSIX access control.



Limitaciones

El máximo número de ACLs efectivos en un fichero o directorio es de 28.

A modo práctico esto significa que como máximo un blob puede ser gobernado por hasta 28 DSAs, si ningún otro sistema aplica ACLs a ese blob.

Al dar permisos a los ficheros particionados, si los ficheros de esas particiones cambian o se añaden más, no contendrán los permisos que tienen otras partes del mismo.

Ejemplo de configuración

```
totplugin:
  location: http://<host>:<port>/plugin/azure/storage/api/v1
  server:
    url: http://<totserver>:<port>/tot/
    keep-alive-seconds: 60
  connection:
    pathSeparator: "/"
    sampleRows: 15
    storageAccount: totpluginazure
    clientId: "6a0de2axxxxxxxx8b5da496f"
    tenantId: "aef41dxxxxxxxxxfb98ef04ded"
    secret: "6gA0Pkxxxxxxxxxu6_O8avrk__h"

server:
```

```
port: 16000
compression:
  enabled: false  # Whether response compression is enabled.
```