Tot plugin GCP 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

Se utilizan los métodos que ofrece el driver de Google mediante el cual se accede a los distintos recursos.

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 bucket.
- **physicalName** y name con el mismo valor, el nombre del blob correspondiente.
- **path** con el path del bucket 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).

Al finalizar el workflow de creación cuando un objeto es gobernado se mandan a Tot todo el metadato disponible para ese objeto, es decir, todos los atributos existentes del objeto creado.

Muestreo de datos

Para realizar el sampleo de datos se hace la consulta al blob correspondiente obteniendo el dato según el contenido especificado en el objeto a samplear y según un número limitado de los mismos especificados por configuración. Se ofusca el contenido de la columna en cuestión si es necesario.

Gobierno activo

Dentro del gobierno activo vamos a tener disponible las siguientes operaciones siempre que se tenga permiso para ello a través del api proporcionado por Google:

- Creación de grupos
- Creación de usuarios
- Borrado de usuarios
- Borrado de un objeto de un grupo

Para ello el plugin va a delegar en otro plugin encargado de la gestión de identidades y accesos el crear estos usuarios y grupos. Una vez creados, el plugin se encargará de dar el acceso adecuado a los usuarios en los recursos solicitados.
Por defecto el plugin está configurado para utilizar el plugin Gcp IAM, pudiendo cambiarlo si fuera necesario poniendo el valor correspondiente en la propiedad imAri:

```
aris:
   - ari: "anja:totplugin:extract:/gcp/storage/<zone>/"
   - ari: "anja:totplugin:sample:/gcp/storage/<zone>/"
   - ari: "anja:totplugin:im:/gcp/storage/<zone>/"
 imAri: "anja:totplugin:im:/gcp/iam/<zone>/"
```

Hay que tener en cuenta que Google impone ciertas restricciones a los nombres de los recursos con los que tratar. Podemos ver las siguientes restricciones:

- Tener entre 1 y 63 caracteres de longitud.
- Cumplir con las convenciones de RFC 1035.
- Coincidir con la expresión regular [a-z]([-a-z0-9]*[a-z0-9])?. Esto significa que el primer carácter debe ser una letra minúscula y todos los caracteres siguientes deben ser guiones, minúsculas o dígitos, excepto el último carácter, que no puede ser un guión.

Para ampliar estas restricciones ir a Naming resources | Google

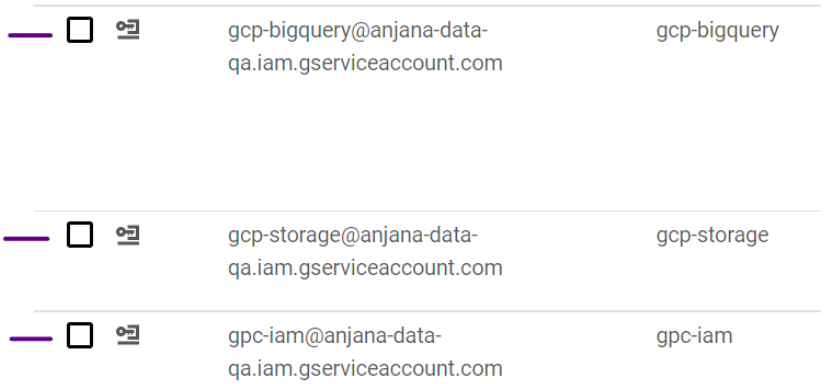
Se pueden aplicar ciertas restricciones a la cantidad de usuarios relacionados con cada recurso. Estas limitaciones se pueden consultar en https://cloud.google.com/iam/quotas

Credenciales requeridas

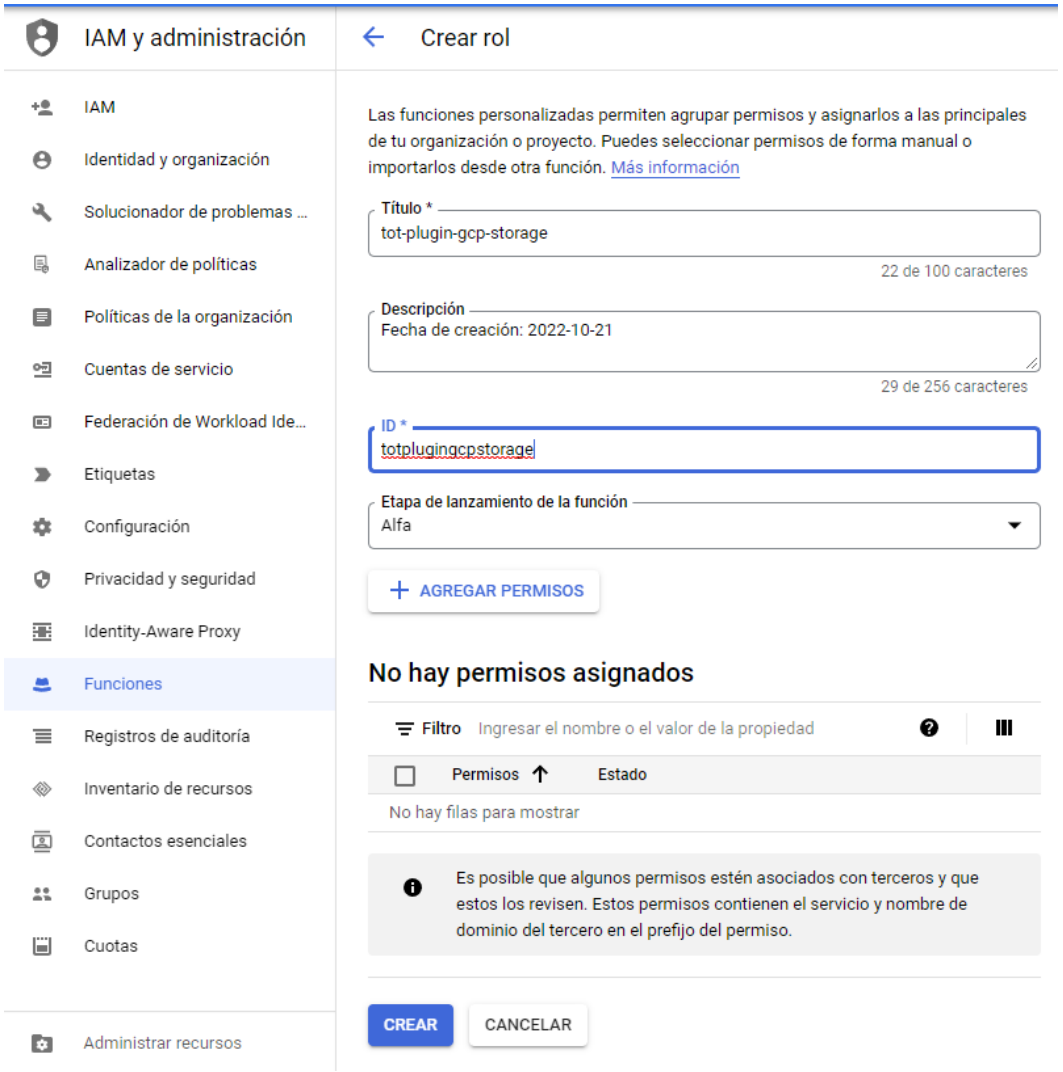
Las credenciales requeridas se deberán configurar en el fichero yaml que utilizaremos para arrancar el servicio en la propiedad "`totplugin.connection.credentialsContent`"

Crear la cuenta de servicio

Para GCP será necesario crear una cuenta de servicio en IAM para cada plugin de forma individual y tras eso asignarle los permisos necesarios para la ejecución de las tareas específicas de cada plugin.



Para personalizar los permisos de forma más acorde será necesaria la creación de roles personalizados en los cuáles se engloban los permisos que luego son asociados a las cuentas de servicio.



Extracción de metadatos

Los permisos utilizados son los siguientes:

- `storage.objects.get`
- `storage.objects.list`

Muestreo de datos

Los permisos utilizados son los siguientes:

- `storage.objects.get`
- `Storage.objects.list`

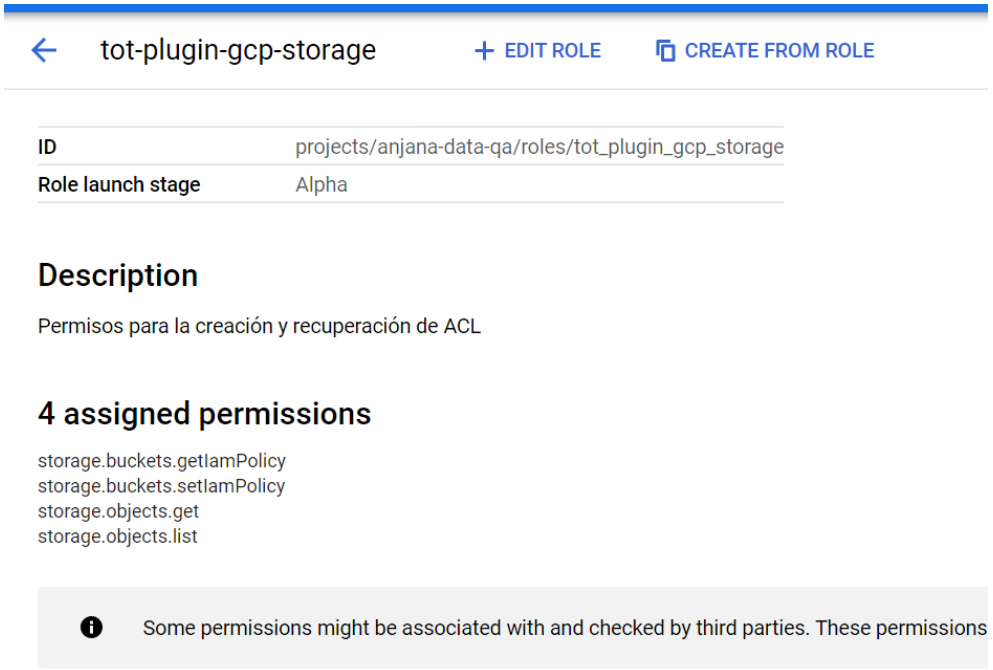
Gobierno activo

La gestión de acceso requiere el plugin “Tot plugin GCP IAM” para que genere los roles(funciones) custom que representan a los DSA.

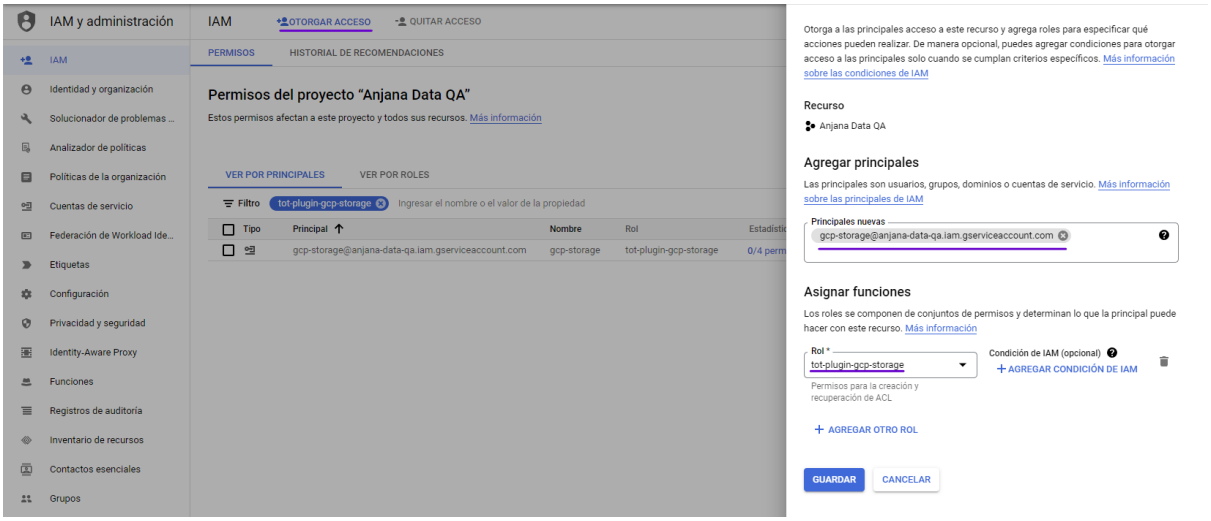
Los permisos utilizados son los siguientes:

- `storage.buckets.getIamPolicy`
- `storage.buckets.setIamPolicy`
- `storage.objects.get`
- `storage.objects.list`

En resumen los permisos utilizados para el rol personalizado serán los siguientes:



Para asignar los permisos a la cuenta de servicio de storage tendremos que:



Ejemplo de configuración

Se han de revisar las configuraciones comunes en el doc de configuraciones “Anjana Data - Microservices configuration”
Configuraciones específicas:

- `connection`:
  - `credentialsContent`: Credenciales de acceso a GCP.
  - `sample-rows`: Tamaño del muestreo de datos.
  - `path-separator`: Separador que GCP usa en los roles

```
server:
  port: 15003

totplugin:
  location: http://totplugingcpstorageserver:15003/plugin/storage/api/v1
  server:
    url: http://totserver:15000/tot/
  aris:
    - ari: "anja:totplugin:extract:/Google/gcpStorage/devQA/"
    - ari: "anja:totplugin:sample:/Google/gcpStorage/devQA/"
    - ari: "anja:totplugin:im:/Google/gcpStorage/devQA/"
      imAri: "anja:totplugin:im:/Google/gcpIam/devQA/"
  connection:
    pathSeparator: "/"
    sampleRows: 5
    bucket: qa43
    credentialsContent: |
      {
        "type": "service_account",
        "project_id": "anjana-data-qa",
        "private_key_id": "**********************",
        "private_key": "-----BEGIN PRIVATE KEY-----\n\n-----END PRIVATE KEY-----\n",
        "client_email": "gpc-iam@*****.com",
        "client_id": "**********************",
        "auth_uri": "https://accounts.google.com/o/oauth2/auth",
        "token_uri": "https://oauth2.googleapis.com/token",
        "auth_provider_x509_cert_url": "https://www.googleapis.com/oauth2/v1/certs",
        "client_x509_cert_url":
"https://www.googleapis.com/robot/v1/metadata/x509/gcp-storage%40*****.com"
      }
```