Tot plugin AWS S3

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

Para el listado de estructuras, se recupera la lista de *buckets* accesibles y todos los objetos con el objetivo de retornar un mapa de contenido. El comportamiento será un poco diferente si se configura un *bucket* en el fichero YML de configuración. En tal caso se limitará el listado al contenido de ese *bucket*.
El valor retornado en el listado de estructuras es un objeto con el nombre de la estructura (que es un concatenado de la infraestructura, la tecnología y la zona) y una lista de objetos que en ella se encuentran. En este caso la estructura se corresponde con el *bucket* y el objeto con el objeto en Amazon S3.
Para la extracción de metadata de un objeto se utilizan las mismas herramientas, se lee el contenido del objeto para extraer su metadato y se devuelve el resultado.
Después de ejecutar la extracción del metadato obtendremos un objeto con un nombre (*elementName*), con una lista de atributos clave valor (como la infraestructura, la tecnología y la zona) y además una lista de campos (*fields*) asociados al mismo, también con su lista de atributos clave valor.
Para una exitosa extracción de metadato, los atributos deben llamarse igual en la tabla *attribute_definition* para el campo *name* con el objetivo de que aparezcan en pantalla:

- ***physicalName:*** nombre con el mismo valor que el objeto.
- ***path:*** con la concatenación de los valores del objeto en Amazon S3.
- ***infrastructure:*** con el valor correspondiente.
- ***technology:*** con el valor correspondiente.
- ***zone:*** con el valor correspondiente.

Para los atributos dentro de los *fields* deberá ocurrir lo mismo:

- ***name:*** nombre del campo correspondiente.
- ***physicalName:*** con el valor del campo correspondiente.
- ***fieldDataType*** con el tipado del campo correspondiente.
- ***position:*** con la posición que ocupa el campo correspondiente.
- ***nullable:*** indicando si el campo correspondiente admite nulos o no.
- ***description:*** con el valor del campo correspondiente.

Tipos soportados

El presente plugin soporta la extracción de metadata de los siguientes tipos de elemento:

- Parquet (“.parquet”)
- Avro (“.avro”)
- Excel (“.xls” , “.xlsx”)
- CSV (“.csv”)

Adicionalmente el plugin tiene la capacidad de reconocer las típicas pilas de ficheros generados desde procesamientos paralelos considerándose un solo bloque de información. El patrón aplicado a dicho reconocimiento es <fichero><secuencia>.<extensión> o <fichero>.<secuencia>.<extensión>
Dependiendo del tipo de elemento seleccionado, nos enviará diferentes atributos relativos a los campos del recurso pedido.

- 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
    - **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
    - **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
    - **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

Para el muestreo de datos, se localiza el objeto a muestrear (hasta el número máximo de resultados configurados), se leen los contenidos de dichos ficheros utilizando librerías de Apache según la tipología de los ficheros y se devuelven los resultados.
El valor retornado en el muestreo de datos es un objeto que contiene cabeceras (*headers*) y valores (*values*).
En las cabeceras (*headers*) se incluyen los nombres de los atributos de los objetos que se deben retornar tras el muestreo de los datos.
En los valores (*values*) se incluyen la lista de valores de cada cabecera para cada objeto que se desea retornar. Esto permitirá que el plugin pueda devolver, aparte de atributos como el nombre del fichero o el contenido del fichero, el resto de datos y metadatos.

Gobierno activo de estructuras

En el protocolo S3 las rutas son emuladas con lo cual no es posible pre-provisionar dichos elementos.

Gobierno activo de accesos

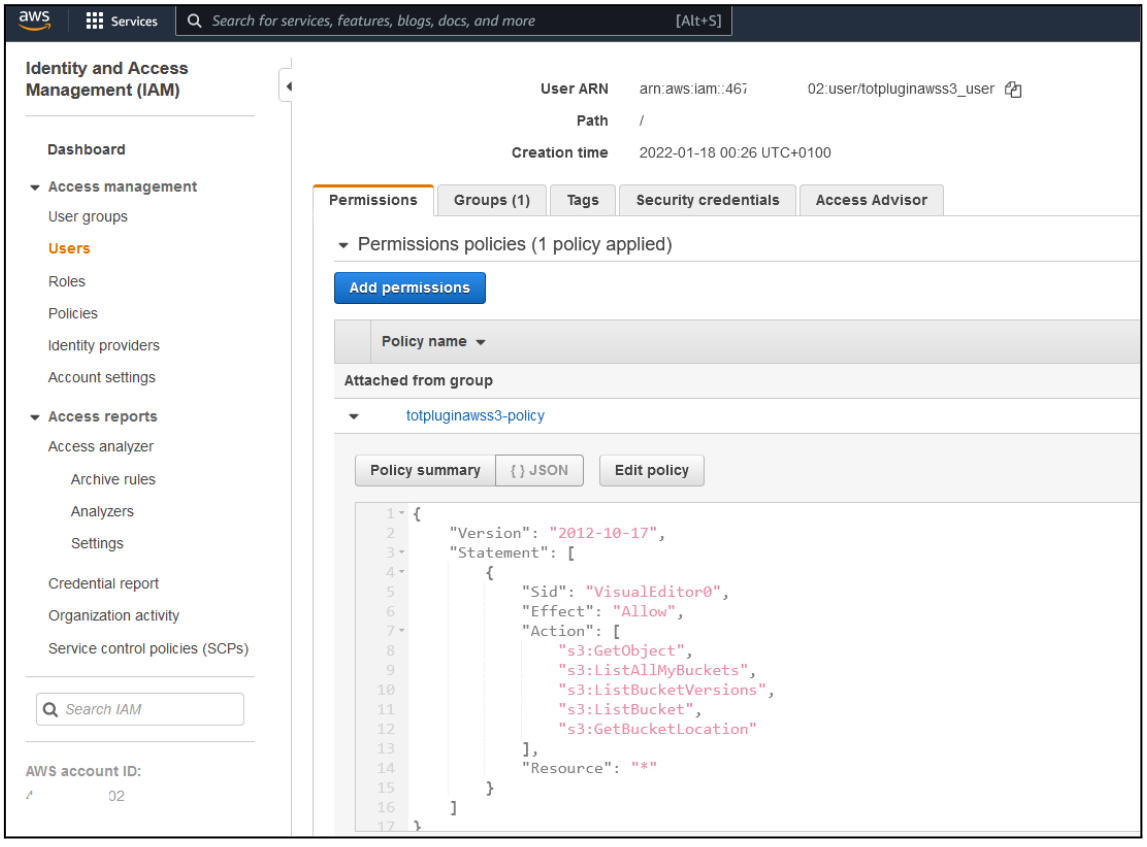
La gestión de accesos sobre esta tecnología se realiza directamente en AWS IAM por lo cual se delega este tipo de actuación en el plugin para dicha tecnología, siendo imprescindible la presencia de este último por tanto para disponer de la funcionalidad.

Credenciales requeridas

Extracción de metadatos

Si se desea realizar acciones de extracción de metadato se requiere de una conexión con Amazon S3. Para establecer dicha conexión se requiere un *accessKey* y un *secretKey* de la cuenta que se proporcione a Anjana para manejar el gobierno del dato. De manera opcional, se requiere de un proxy si el usuario no quiere que se conecte directamente a Amazon S3.
De los datos de Amazon S3 se requiere la región en la que se encuentra y de manera opcional, si sólo se quiere gobernar un único *bucket*, el *bucket* deseado (si no se incluye un *bucket* se gobiernan todos los *buckets* en los que la cuenta tenga permisos).
Por su parte, Amazon S3 define un conjunto de acciones[1] que se pueden especificar en una política y que ayudarán a que el usuario con acceso a la tecnología pueda obtener la información deseada. Para este plugin resultan interesantes los siguientes acciones:

- **s3:ListAllMyBuckets** para listar todos los *buckets* del usuario autenticado.
- **s3:ListBucket** para listar el contenido de un *bucket*.
- **s3:GetBucketLocation** para retornar la región en la que reside el *bucket*.
- **s3:GetObject** para retornar objetos de Amazon S3. Para poder leer el objeto se deberá, además, tener permisos de lectura sobre el mismo.



[1] Acciones permitidas en Amazon S3: https://docs.aws.amazon.com/AmazonS3/latest/API/API_Operations_Amazon_Simple_Storage_Service.html

Muestreo de datos

Para desencadenar acciones relacionadas con el muestreo de datos se requiere de la misma configuración y credenciales que las mencionadas anteriormente para la extracción de metadatos (apartado 2.2).

Gobierno activo de estructuras

En el protocolo S3 las rutas son emuladas con lo cual no es posible pre-provisionar dichos elementos.

Gobierno activo de accesos

La gestión de accesos sobre esta tecnología se realiza directamente en AWS IAM por lo cual se delega este tipo de actuación en el plugin para dicha tecnología, siendo imprescindible la presencia de este último por tanto para disponer de la funcionalidad.

Configuración necesaria

```yaml
server:
 port: 15007

totplugin:
 location: http://totpluginawss3server:15007/plugin/aws/s3/api/v1
 server:
   url: http://totserver:15000/tot/
 aris:
   - ari: "anja:totplugin:extract:/AWS/S3/DEV/"
   - ari: "anja:totplugin:sample:/AWS/S3/DEV/"
   - ari: "anja:totplugin:im:/AWS/S3/DEV/"
     imAri: "anja:totplugin:im:/AWS/S3/IAM/"
 connection:
   #proxy:
   accessKey: ""
   secretKey: ""
   bucket: ""
   region: ""
```

Server:

- port: El puerto en el que se va a desplegar el plugin.

TotPlugin (apartado con la configuración específica del plugin):

- Location: URL del plugin cuando está desplegado (lo que se debe modificar es el host y el puerto, la ruta de entrada no debe modificarse)
- Server:
  - Url: URL de tot
- Aris:

- ari: ARI usada para registrarse en Tot y poder ser referido y llamado según eventos en Anjana.

- Connection (apartado con la configuración relativa a las credenciales de conexión con Amazon S3):
  - proxy(opcional): La url del proxy si no se quiere que el plugin tenga conexión directa con S3
  - accessKey: Clave de acceso de la cuenta de Anjana generada para el plugin.
  - secretKey: La contraseña de acceso de la cuenta de Anjana generada para el plugin.
  - bucket(opcional): Si se define, limita a gobernar un bucket específico. En caso de que no se defina, se gobernaran todos los buckets accesible desde la cuenta.
  - region: La región donde se encuentra registrado el S3.