Artículo
Ricardo Paiva · 29 ene, 2021 Lectura de 8 min

Open Exchange

# Cómo leer los datos de AWS S3 sobre el COVID como tablas SQL en IRIS

¡Hola Desarroladores!

IRIS External Table es un proyecto de código abierto de la comunidad de InterSystems, que permite utilizar archivos almacenados en el sistema de archivos local y almacenar objetos en la nube como AWS S3 y tablas SQL.

Se puede encontrar en GitHub https://github.com/intersystems-community/IRIS-ExternalTable Open Exchange https://openexchange.intersystems.com/package/IRIS-External-Table y está incluido en el administrador de paquetes InterSystems Package Manager (ZPM).

Para instalar External Table desde GitHub, utilice:

```
git clone https://github.com/antonum/IRIS-ExternalTable.git
iris session iris
USER>set sc = ##class(%SYSTEM.OBJ).LoadDir("<path-to>/IRIS-ExternalTable/src", "ck",,1)
```

Para instalarlo con el ZPM Package Manager, utilice:

```
USER>zpm "install external-table"
```

## Cómo trabajar con archivos locales

Crearemos un archivo simple que tiene este aspecto:

```
a1,b1
a2,b2
```

Abra su editor favorito y cree el archivo o utilice solo una línea de comandos en Linux/Mac:

```
echo $'a1,b1\na2,b2' > /tmp/test.txt
```

Cree una tabla SQL en IRIS para representar este archivo:

```
create table test (col1 char(10),col2 char(10))
```

Convierta la tabla para utilizar el almacenamiento externo:

```
CALL EXT.ConvertToExternal(
    'test',
    '{
        "adapter":"EXT.LocalFile",
        "location":"/tmp/test.txt",
        "delimiter": ","
    }')
```

Y finalmente, consulte la tabla:

```
select * from test
```

Si todo funciona según lo previsto, debería ver el resultado de la siguiente forma:

```
col1    col2
a1  b1
a2  b2
```

Ahora regrese al editor, modifique el contenido del archivo y ejecute nuevamente la consulta SQL. ¡¡¡Tarán!!! Está leyendo nuevos valores de su archivo local en SQL.

```
col1    col2
a1  b1
a2  b99
```

## Cómo leer los datos desde S3

En https://covid19-lake.s3.amazonaws.com/index.html puede acceder a los datos de la COVID que se actualizan constantemente, estos se almacenan por AWS en el lago de datos públicos.

Intentaremos acceder a una de las fuentes de datos en este lago de datos: s3://covid19-lake/rearc-covid-19-nyt-data-in-usa/csv/us-states

Si tiene instalada la herramienta de línea de comandos para AWS, puede repetir los siguientes pasos. Si no es así, vaya directamente a la parte de SQL. No es necesario que tenga ningún componente específico de AWS instalado en su equipo para continuar con la parte de SQL.

```
$ aws s3 ls s3://covid19-lake/rearc-covid-19-nyt-data-in-usa/csv/us-states/
2020-12-04 17:19:10     510572 us-states.csv

$ aws s3 cp s3://covid19-lake/rearc-covid-19-nyt-data-in-usa/csv/us-states/us-
states.csv .
download: s3://covid19-lake/rearc-covid-19-nyt-data-in-usa/csv/us-states/us-
states.csv to ./us-states.csv

$ head us-states.csv
date,state,fips,cases,deaths
2020-01-21,Washington,53,1,0
2020-01-22,Washington,53,1,0
```

```
2020-01-23,Washington,53,1,0
2020-01-24,Illinois,17,1,0
2020-01-24,Washington,53,1,0
2020-01-25,California,06,1,0
2020-01-25,Illinois,17,1,0
2020-01-25,Washington,53,1,0
2020-01-26,Arizona,04,1,0
```

Por lo tanto, tenemos un archivo con una estructura bastante simple y cinco campos delimitados.

Para mostrar esta carpeta S3 como en External Table, primero necesitamos crear una tabla " regular " con la estructura deseada:

```
-- create external table
create table covid_by_state (
    "date" DATE,
    "state" VARCHAR(20),
    fips INT,
    cases INT,
    deaths INT
)
```

Tenga en cuenta que algunos campos de datos como " Date " son palabras reservadas en el SQL de IRIS y deben escribirse entre comillas dobles. Entonces, necesitamos convertir esta tabla " regular " en la tabla " externa " , basada en el bucket AWS S3 y con el tipo CSV.

```
 -- convert table to external storage
call EXT.ConvertToExternal(
    'covid_by_state',
    '{
    "adapter":"EXT.AWSS3",
    "location":"s3://covid19-lake/rearc-covid-19-nyt-data-in-usa/csv/us-states/",
    "type": "csv",
    "delimiter": ",",
    "skipHeaders": 1
    }'
)
```

Si observa detenidamente, en EXT.ExternalTable los argumentos de los procedimientos son el nombre de la tabla y luego la cadena JSON, además contiene varios parámetros como la ubicación para buscar archivos, el adaptador para utilizarlos, un delimitador, etc. Además, External Table de AWS S3 es compatible con el almacenamiento de Azure BLOB, Google Cloud Buckets y el sistema de archivos local. El repositorio de GitHub contiene referencias para la sintaxis y opciones que son compatibles con todos los formatos.

Y finalmente, consulte la tabla:

```
-- query the table
select top 10 * from covid_by_state order by "date" desc

[SQL]USER>>select top 10 * from covid_by_state order by "date" desc
2.  select top 10 * from covid_by_state order by "date" desc

date    state   fips    cases   deaths
```

```
2020-12-06  Alabama 01  269877  3889
2020-12-06  Alaska  02  36847   136
2020-12-06  Arizona 04  364276  6950
2020-12-06  Arkansas    05  170924  2660
2020-12-06  California  06  1371940 19937
2020-12-06  Colorado    08  262460  3437
2020-12-06  Connecticut 09  127715  5146
2020-12-06  Delaware    10  39912   793
2020-12-06  District of Columbia    11  23136   697
2020-12-06  Florida 12  1058066 19176
```

Es comprensible que se necesite más tiempo para consultar los datos de la tabla remota, que para consultar la tabla " nativa de IRIS " o la tabla basada en el global, pero esta se almacena y actualiza completamente en la nube, y en segundo plano se extrae a IRIS.

Exploremos un par de funciones adicionales de External Table.

## %PATH y las tablas basadas en varios archivos

La carpeta de nuestro ejemplo, que se encuentra en el bucket, contiene solo un archivo. Lo más común es que tenga varios archivos de la misma estructura, donde el nombre del archivo identifique tanto al registro de la hora como al identificador de algún otro atributo que queramos utilizar en nuestras consultas.

El campo %PATH se agrega automáticamente a cada tabla externa y contiene la ruta completa hacia el archivo de donde se recuperó la fila.

```
select top 5 %PATH,* from covid_by_state
```

```
%PATH   date    state   fips    cases   deaths
s3://covid19-lake/rearc-covid-19-nyt-data-in-usa/csv/us-states/us-
states.csv    2020-01-21  Washington  53  1   0
s3://covid19-lake/rearc-covid-19-nyt-data-in-usa/csv/us-states/us-
states.csv    2020-01-22  Washington  53  1   0
s3://covid19-lake/rearc-covid-19-nyt-data-in-usa/csv/us-states/us-
states.csv    2020-01-23  Washington  53  1   0
s3://covid19-lake/rearc-covid-19-nyt-data-in-usa/csv/us-states/us-
states.csv    2020-01-24  Illinois    17  1   0
s3://covid19-lake/rearc-covid-19-nyt-data-in-usa/csv/us-states/us-
states.csv    2020-01-24  Washington  53  1   0
```

Puede utilizar el campo %PATH en sus consultas SQL como cualquier otro campo.

## De datos ETL a " tablas regulares "

Si su tarea es cargar datos de S3 en una tabla IRIS, puede utilizar External Table como una herramienta ETL. Simplemente haga lo siguiente:

```
INSERT INTO internal_table SELECT * FROM external_table
```

En nuestro caso, si queremos copiar los datos COVID de S3 a la tabla local:

```
--create local table
create table covid_by_state_local (
    "date" DATE,
    "state" VARCHAR(100),
    fips INT,
    cases INT,
    deaths INT
)
--ETL from External to Local table
INSERT INTO covid_by_state_local SELECT TO_DATE("date",'YYYY-MM-
DD'),state,fips,cases,deaths FROM covid_by_state
```

## UNIÓN entre IRIS, tablas nativas y externas. Consultas federadas

External Table es una tabla SQL. Se puede unir con otras tablas, utilizarse en subconsultas y sistemas de archivos tipo UNION. Incluso puede combinar la tabla " Regular " de IRIS y dos o más tablas externas que provengan de diferentes fuentes en la misma consulta SQL.

Intente crear una tabla regular, por ejemplo, haga coincidir los nombres de los estados con sus códigos como en el caso de Washington y WA. Y únalos con nuestra tabla basada en S3.

```
create table state_codes (name varchar(100), code char(2))
insert into state_codes values ('Washington','WA')
insert into state_codes values ('Illinois','IL')

select top 10 "date", state, code, cases from covid_by_state join state_codes on stat
e=name
```

Cambie " join " por " left join " para incluir aquellas filas donde el código del estado no esté definido. Como puede ver, el resultado es una combinación de datos provenientes de S3 y su tabla nativa de IRIS.

## Acceso seguro a la información

El lago de datos Covid en AWS es público. Cualquier persona puede leer los datos que provengan de esta fuente sin la necesidad de tener alguna autenticación o autorización. En la vida real seguramente quiere acceder a sus datos de una forma segura, donde se evite que extraños echen un vistazo a sus archivos. Los detalles completos sobre AWS Identity y Access Management (IAM) están fuera del alcance de este artículo. Pero lo mínimo que debe saber es que necesita por lo menos la clave de acceso a la cuenta y la información confidencial de AWS para acceder a los datos privados de su cuenta.

AWS utiliza la autenticación de claves/información confidencial de la cuenta para firmar las solicitudes.
https://docs.aws.amazon.com/general/latest/gr/aws-sec-cred-types.html#ac...

Si está ejecutando IRIS External Table en una instancia de EC2, la forma recomendada de lidiar con la autenticación es utilizando las funciones que se encuentran en la instancia de EC2 https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/iam-roles-for-amazon.... De este modo, IRIS External Table podría utilizar los permisos de esa función. No se requiere ninguna configuración adicional.

En una instancia local o que no sea de EC2 es necesario especificar AWSACCESSKEYID y AWSSECRETACCESSKEY, ya sea con la ayuda de variables de entorno o mediante la instalación y configuración del cliente CLI de AWS.

```
export AWS_ACCESS_KEY_ID=AKIAEXAMPLEKEY
export AWS_SECRET_ACCESS_KEY=111222333abcdefghigklmnopqrst
```

Asegúrese de que la variable de entorno sea visible dentro del proceso de IRIS. Puede verificarlo al ejecutar:

```
USER>write $system.Util.GetEnviron("AWS_ACCESS_KEY_ID")
```

Esto debería emitir el valor de la clave.

O instale el CLI de AWS, mediante instrucciones que se encuentran aquí: https://docs.aws.amazon.com/cli/latest/userguide/install-cliv2-linux.html y ejecutar:

```
aws configure
```

Entonces External Table podrá leer las credenciales desde los archivos de configuración para el CLI de AWS. Posiblemente su shell interactivo y el proceso de IRIS estén ejecutándose en cuentas diferentes, asegúrese de ejecutar aws configure con la misma cuenta que su proceso de IRIS.

#Analítica #CSV #Interoperabilidad #Mejores prácticas #Nube #SQL #InterSystems IRIS
Ir a la aplicación en InterSystems Open Exchange