Amazon S3 Connector

Plugin module	Data provider name	Automatic metadata detection
AWS	S3	No

The Amazon S3 connector allows external tables to fetch data from Amazon's S3 cloud storage service. It is compatible with Avro, JSON, and delimited text files such as CSV. See the reference sheets about those formats for specific help.

Prerequisites

- Kognitio version 8.1 or later with 64-bit `libcurl`, `libxml2`, `libssl` and `libcrypto` installed on all DB nodes.
- Amazon Web Services account with access key ID and secret key ID, and read access to an S3 storage bucket; alternatively, for Kognitio systems of version 8.2.1 or later running in AWS, the DB nodes may have an IAM Role giving them S3 read permission on the bucket.
- HTTPS access to `s3.amazonaws.com`, optionally via a proxy.

Examples

Load the plugin

```
create module aws;
alter module aws set mode active;
```

Create a connector

```
create connector mys3bucket source s3
target 'uri_location s3://nameofyourbucket,
accesskey AKIAIOSFODNN7EXAMPLE,
secretkey wJalrXUtnFEMIK7MDENGbPxRfiCYEXAMPLEKEY';
```

List objects matching wildcard

```
external table from mys3bucket target 'list items/*.csv';
```

Create an external table from a list of files in S3

```
create external table ext_customer (
        id int,
        name varchar(100),
        company varchar(100),
        address varchar(400),
        countrycode char(3)
) from mys3bucket
target 'uri_path "customers/jan2013.csv
customers/feb2013.csv
customers/mar2013.csv
customers/old/*201[012].csv"';
```

Attributes

Attribute name	Type	Default	Description
`uri_location`	string	none (required)	The Amazon S3 bucket in which to find the files, in the form `s3://bucketname`
`file` or `uri_path`	string	none (required)	Object name, or directory, or bash-style wildcard pattern, e.g. `sales/2013/*.csv`. May be a space-separated list of patterns.
`accesskey`	string	none (required if no IAM Role)	AWS access key ID. The query above contains a non-working example.
`secretkey`	string		AWS secret key ID. The query above contains a non-working example.
`iamrole`	string	IAM Role of the DB node executing the request	Obtain temporary credentials using this AWS IAM role. If given, this must match the role the DB node is in. If this is not given, but the node is in an IAM Role, we default to obtaining credentials using that role.
`region` (v8.2 onwards)	string	none	Your bucket's region, e.g. `eu-west-1`. Set this if you use Kognitio 8.2.3 or earlier and need to use AWS Signature v4.
`server`	string	s3.amazonaws.com	S3 server endpoint. If region is specified, default is s3-*region*.amazonaws.com.
`proxy`	string	no proxy	Address and port number of web proxy to use, e.g. `proxy:8080`
`proxyuser`/`proxypass`	string		User/password to use to authenticate with the web proxy, if required
`certfile`/`certpath`	string	libcurl default	Specify CA certificate file, or directory to look in for cert files, for libcurl.
`wholefiles`	boolean	false	If true, don't divide files into chunks for processing by separate threads
`requesterpays`	boolean	false	Accept that the requester, not the bucket owner, is billed for the data transfer. This is required to access requester-pays S3 buckets.
`list`	string	none	Don't fetch files; instead list files matching this wildcard pattern. If this option is given, any `file` or `uri_path` attribute is ignored.

Additional attributes can be used to specify how the files are formatted; see the **Target String Format Attributes** reference sheet.

Notes

- If any object names end with `.gz` they are assumed to be gzip files and the plugin decompresses them automatically. Compressed files are not divided into blocks; one thread fetches the whole file.
- If you are running in AWS on version 8.2.1 or later, and all the DB nodes in the system have an IAM Role which has read access to the S3 bucket, and you want the connector to obtain temporary credentials using this role, you do not need to specify `accesskey` and `secretkey`.