Loading Avro files using connectors

Plugin modules	Data provider names	Automatic metadata detection
`HADOOP, AWS, OSFILE`	`HDFS, MAPRFS, S3, OSFILE`	Yes

This reference sheet is about creating external tables which are backed by Avro files in HDFS, MapR-FS, S3 or the OS filesystem. In the simplest case, all you have to do is set the attribute `fmt_avro` to 1, and set the `file` attribute to point to your Avro files. The connector will do the rest.

Prerequisites

- Kognitio server version 8.2.1 or later.
- You must be using one of the block-loading connectors, which are HDFS, MAPRFS, S3 and OSFILE.
- All Avro files referenced by an external table must have an identical schema.

Examples

Create an external table

```
create external table ext_items
from {connector name}
target '
    fmt_avro 1,
    file /user/joe/data/items.avro
';
```

The table column types will be looked up from the Avro schema in the Avro file. If you want to select only a subset of the fields, you'll have to specify a format string (see below).

Create external table with a format string

```
create external table ext_items (
       id int,
       name varchar(100),
       price decimal(7,2)
) from {connector name}
target 'fmt_avro 1,
        file /user/joe/data/items.avro,
        fmt_avro_project "itemid, itemname,
                          itemprice"
';
```

Attributes

Attribute name	Type	Default	Description
`fmt_avro`	boolean	false	If set, it tells the connector that the files are in Avro format.
`fmt_avro_project`	string	none	Comma-separated list of Avro fullnames indicating which fields to project.
`fmt_avro_json_format`	boolean	none	If set, each Avro object will be treated like the equivalent JSON object, and the value of the `fmt_avro_project` attribute is expected to be a JSON format string, which will be used to project the columns. See the *Loading JSON* reference sheet for more details.

Notes

Each file matched by the value of the `file` attribute must be an Avro *Object Container File*, defined here: https://avro.apache.org/docs/1.8.2/spec.html#Object+Container+Files

All Avro files matched by the value of the `file` attribute must contain an identical Avro schema. If this is not the case, attempting to create the external table will give an error. If after the external table is created, new files are introduced into the filesystem which match any wildcard pattern given in the `file` attribute and whose schema does not match the other files, behaviour is undefined.

If some of the fields appear in the path name, like this:

```
/user/joe/data/item_manufacturer_id=10/items.avro
/user/joe/data/item_manufacturer_id=20/items.avro
```

Then in the Avro project string (`fmt_avro_project`), use `inputfilename("<name>")` to get at the value for that name.

```
create external table ext_items (
       id int, manufacturer_id int, name varchar(100), price decimal(7,2)
) from {connector name}
target 'fmt_avro 1, file /user/joe/data/*/items.avro,
        fmt_avro_project "itemid, inputfilename(\"item_manufacturer_id\"), itemname, itemprice"
';
```