Representation of VOTable as Parquet

Status update

Brigitta Sipőcz (Caltech/IPAC-IRSA) with Andreas Faisst and Gregory Dubois-Felsmann

as part of the Fornax Project

Motivation

See previous, HiPSCat talk

Large catalogs distributed in the cloud as Parquet

- IRSA catalogs in the cloud: https://irsa.ipac.caltech.edu/cloud_access/
  - AllWISE Source Catalog
  - NEOWISE-R Single-exposure Source Table

All-sky bulk data access for e.g cross-match

Using these catalogs should fit nicely into existing user workflows

Existing tools should “just work” – e.g. Parquet tools and astropy

Future options to return parquet with VOTable metadata from services

VOTable as Parquet metadata

New astropy I/O formats:

- `'votable.parquet'`
  - New in astropy 6.0 (Nov 2023 release)
  - VOTable with a separate parquet file for data, similar to FITS serialization

VOTable as Parquet metadata

New astropy I/O formats:

- `'votable.parquet'`

```
ids = [f"COSMOS_{ii:03g}" for ii in range(number_of_objects)]
mass = np.random.uniform(low=1e8, high=1e10, size=number_of_objects)
sfr = np.random.uniform(low=1, high=100, size=number_of_objects)
redshift = np.random.uniform(low=0, high=3, size=number_of_objects)
input_table = Table([ids, redshift, mass, sfr], names=["id", "z", "mass", "sfr"])

input_table.write('demo.vot', format='votable.parquet', column_metadata=column_metadata)

Table.read('demo.vot')
```

VOTable as Parquet metadata

```
<?xml version="1.0" encoding="utf-8"?>
<!-- Produced with astropy.io.votable version 7.0.0.dev92+gecfcf0f0c7.d20240519
     http://www.astropy.org/ -->
<VOTABLE version="1.4" xmlns="http://www.ivoa.net/xml/VOTable/v1.3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-i
http://www.ivoa.net/xml/VOTable/v1.3 http://www.ivoa.net/xml/VOTable/VOTable-1.4.xsd">
 <RESOURCE type="results">
  <TABLE>
   <FIELD ID="id" arraysize="10" datatype="unicodeChar" name="id" ucd="meta.id" unit="---" utype="none"/>
   <FIELD ID="z" datatype="double" name="z" ucd="src.redshift" unit="---" utype="none"/>
   <FIELD ID="mass" datatype="double" name="mass" ucd="phys.mass" unit="solMass" utype="none"/>
   <FIELD ID="sfr" datatype="double" name="sfr" ucd="phys.SFR" unit="solMass.yr-1" utype="none"/>
   <DATA>
<PARQUET type="VOTable-remote-file">
<STREAM href="file:demo.vot.parquet"/>
</PARQUET>
   </DATA>
  </TABLE>
 </RESOURCE>
</VOTABLE>
```

VOTable as Parquet metadata

New astropy I/O formats:

- `'votable.parquet'`
  - New in astropy 6.0 (Nov 2023 release)
  - VOTable with a separate parquet file for data, similar to FITS serialization
- `'parquet.votable'`
  - In PR for astropy 7.0 (https://github.com/astropy/astropy/pull/16375)
  - parquet file with VOTable metadata
  - relies on parquet I/O libraries (PyArrow)

VOTable as Parquet metadata

New astropy I/O formats:

- `'parquet.votable'`

```
ids = [f"COSMOS_{ii:03g}" for ii in range(number_of_objects)]
mass = np.random.uniform(low=1e8, high=1e10, size=number_of_objects)
sfr = np.random.uniform(low=1, high=100, size=number_of_objects)
redshift = np.random.uniform(low=0, high=3, size=number_of_objects)
input_table = Table([ids, redshift, mass, sfr], names=["id", "z", "mass", "sfr"])

input_table.write('demo.parquet', format='parquet.votable', metadata=column_metadata)

Table.read('demo.parquet', format='parquet.votable')
```

VOTable as Parquet metadata


```
{b'ARROW:schema': b'/////1AFAAAQAAAAAAAKAA4ABgAFAAgACgAAAAABBAAQAAAAAAAKAAwAAAAEAAgACgAAAFgEAAAEAAAABAAAABQEAADE
BFAwAAPD94bWwgdmVyc2lvbj0nMS4wJyBlbmNvZGluZz0ndXRmLTgnPz4KPG5zMDpWT1RBQkxFIHhtbG5zOm5zMD0iaHR0cDovL3d3dy5pdm9hLm
bWxuczp4c2k9Imh0dHA6Ly93d3cudzMub3JnLzIwMDEvWE1MU2NoZW1hLWluc3RhbmNlIiB2ZXJzaW9uPSIxLjQiIHhzaTpzY2hlbWFMb2NhdGlv
94bWwvVk9UYWJsZS92MS4zIGh0dHA6Ly93d3cuaXZvYS5uZXQveG1sL1ZPVGFibGUvVk9UYWJsZS0xLjQueHNkIj4KIDxuczA6UkVTT1VSQ0UgdH
QUJMRT4KICAgPG5zMDpGSUVMRCBJRD0iaWQiIGFycmF5c2l6ZT0iOCIgZGF0YXR5cGU9InVuaWNvZGVDaGFyIiBuYW1lPSJpZCIgdWNkPSJtZXRh
Jub25lIiAvPgogICA8bnMwOkZJRUxEIEIElEPSJ6IiBkYXRhdHlwZT0iZG91YmxlIiBuYW1lPSJ6IiB1Y2Q9InNyYy5yZWRzaGlmdCIgdW5pdD0iLS
PG5zMDpGSUVMRCBJRD0ibWFzcyIgZGF0YXR5cGU9ImRvdWJsZSIgbmFtZT0ibWFzcyIgdWNkPSJwaHlzLm1hc3MiIHVuaXQ9InNvbE1hc3MiIHV0
ZJRUxEIElElEPSJzZnIiIGRhdGF0eXBlPSJkb3VibGUiIG5hbWU9InNmciIgdWNkPSJwaHlzLlNGUiIgdW5pdD0ic29sTWFzcy55ci0xIiB1dHlwZT
PjxuczA6UEFSUVVFVCB0eXBlPSJQYXJxdWV0LWxvY2FsLVhNTCIgLz48L25zMDpEQVRBPjwvbnMwOlRBQkxFPgogPC9uczA6UkVTT1VSQ0U+Cjwv
EuVk9UYWJsZS1QYXJxdWV0LmNvbnRlbnQAAAAAgP///xQAAAAEAAAABQAAAHV0Zi04AAAAHQAAAElWT0EuVk9UYWJsZS1QYXJxdWV0LmVuY29kaW5
cnF1ZXQtbG9jYWwtWE1MAAAAGQAAAElWT0EuVk9UYWJsZS1QYXJxdWV0LnR5cGUAAAAIAAwABAAIAAgAAAAQAAAABAAAAAMAAAAxLjAAHAAAAElW
Npb24AAAAABAAAAKAAAABgAAAAMAAAAQAAACA////AAABAxAAAAAUAAAABAAAAAAAAAADAAAAc2ZyAK7///8AAAIAqP///wAAAQMQAAAAGAAAAA
/wAAAgDU////AAABAxAAAAAYAAAABAAAAAAAAAABAAAAegAGAAgABgAGAAAAAAACABAAFAAIAAYABwAMAAAAEAAQAAAAAAABBRAAAAAYAAAABAAA
 b'IVOA.VOTable-Parquet.content': b'<?xml version=\'1.0\' encoding=\'utf-8\'?>\n<ns0:VOTABLE xmlns:ns0="http://ww
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" version="1.4" xsi:schemaLocation="http://www.ivoa.net/xml/V
et/xml/VOTable/VOTable-1.4.xsd">\n <ns0:RESOURCE type="results">\n  <ns0:TABLE>\n   <ns0:FIELD ID="id" arraysize=
me="id" ucd="meta.id" unit="---" utype="none" />\n   <ns0:FIELD ID="z" datatype="double" name="z" ucd="src.redsh
>\n   <ns0:FIELD ID="mass" datatype="double" name="mass" ucd="phys.mass" unit="solMass" utype="none" />\n   <ns0
le" name="sfr" ucd="phys.SFR" unit="solMass.yr-1" utype="none" />\n   <ns0:DATA><ns0:PARQUET type="Parquet-local-
>\n </ns0:RESOURCE>\n</ns0:VOTABLE>',
 b'IVOA.VOTable-Parquet.encoding': b'utf-8',
 b'IVOA.VOTable-Parquet.type': b'Parquet-local-XML',
 b'IVOA.VOTable-Parquet.version': b'1.0'}
```

What are we still working on

Finalizing API

- How to handle existing metadata

Collecting more user cases

- astropy Table → parquet → astropy Table
- VOTable → astropy Table → parquet
- parquet → astropy Table → …
- …

Next steps

Community standardization of a Parquet-based data access

More applications to utilize understand VOTable as Parquet