

An ODBC Driver for HDF5

At The HDF Group, we have begun the development of an ODBC driver for HDF5. We would like to use this opportunity to give you an overview of what's coming and invite you to comment.

Open Database Connectivity (ODBC)

ODBC is a standard SQL application programming interface (API) for accessing relational and non-relational data stores. Introduced by Microsoft in 1991, it quickly gained popularity beyond the Windows platform and is today one of the most widely supported APIs for data import and export.

A Use Case

Have you ever tried importing data from an HDF5 file into an Excel spreadsheet, or to dump data from a relational table into an HDF5 file? These are pretty common and straightforward tasks, but none of the choices available today are particularly appealing: Other than ad hoc coding, any solution would involve several auxiliary tools for data extraction, transformation, and loading (ETL). More advanced users may resort to extreme means such as the Microsoft Office COM interface.

Looking at this and other use cases, it is pretty clear that what's been missing all along is a *standard connector* that links HDF5 data containers to clients, from desktop applications to servers. There are several standards and technologies to choose from when building such a connector, but it appears that ODBC is a front-runner.

As an HDF5 user, you may feel some discomfort in the idea of looking at an HDF5 container through "SQL glasses". You may ask, for example, "What happens to all the HDF5 groups? Isn't a SQL database just a flat assortment of tables?" In the next section, we will describe the look-and-feel of the solution we are developing.

The Solution

The solution, as far as you, Excel, or any other application is concerned, is pretty simple: you create (or reuse) an ODBC data source, open it in Excel, select the data you are interested in, and take a sip of coffee while it's loading.

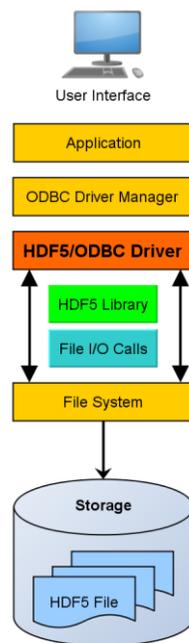
What is the role of that ODBC data source? How does it know which HDF5 file to read from and which dataset to read?

A conversation with an ODBC data source, a so-called *session*, begins with connecting to the data store described as part of the data source's *connection string*. With HDF5, this is simply another way of saying that we are opening a particular HDF5 file. The ODBC data source may contain specific information about which HDF5 dataset(s) the data is supposed to come from, but generally, one will take advantage of the ODBC driver's *catalog functions* to dynamically select the dataset (table), the columns, and the elements (rows) to read, via a SQL WHERE clause.

Although the ODBC driver for HDF5 includes reasonable defaults for mappings between HDF5 types and SQL types, users may have very different and specific ideas about that mapping. In this and similar situations, users can provide an XML descriptor file that, very much like many existing ETL tools, declares mappings from HDF5 datasets to ODBC tables. Such an XML descriptor is then referenced in a data source's connection string and different XML descriptors can be used as different ODBC *views* of the same underlying HDF file.

We Need Your Input!

If the ODBC Driver for HDF5 sounds like a valuable tool for your work, or if you have additional questions about it, turn the page to learn how to contact us and provide input that will help us develop this important solution.



Instructions: [Take the survey](#) or email your answers to the questions below to odbcsurvey@hdfgroup.org.

User Interest Survey -- ODBC Driver for HDF5

1. What is your primary role?
 - a. Analyst
 - b. Developer
 - c. Educator
 - d. Engineer
 - e. Manager
 - f. Researcher
 - g. Other: _____
2. Rate your own ODBC experience/expertise on a 1 – 5 scale (1 – low, 5 – high).
3. Rate your own HDF5 experience/expertise on a 1 – 5 scale (1 – low, 5 – high).
4. Check the top 3 applications with which you would use this driver.
 - a. ArcGIS
 - b. IDL
 - c. Labview
 - d. MATLAB
 - e. MS Office (Excel, Access)
 - f. QlikView
 - g. R
 - h. Relational Database
 - i. SAS
 - j. SPSS
 - k. Others: _____
5. What are the most important benefits of this driver (other than “being able to use applications in the previous question”)?
6. In your estimation, what would be the approximate time or costs savings of using this ODBC driver?
7. On which platforms would you use this ODBC driver? (This is the system that runs your client application and has access to your HDF5 files.)
 - a. AIX
 - b. HP-UX
 - c. Linux
 - d. Mac OS X
 - e. Solaris
 - f. VMS
 - g. Windows
 - h. Others: _____
 - i. RedHat
 - ii. SUSE
8. Do you have any special expectations for this ODBC driver beyond the features discussed on the data sheet?
9. Would you be willing to participate in a beta program?
10. Other comments / questions / concerns?



1800 South Oak St., Suite 203, Champaign, Illinois 61820
(217) 531-6100
www.hdfgroup.org
Email: odbcsurvey@hdfgroup.org