Contribution ID: **94** Type: **not specified**

Finding Needles in a Huge DataStack

Monday, July 3, 2006 5:10 PM (30 minutes)

Many tools exist in the Python world to handle persistent data. Most of them are high-level wrappers to access well-known relational databases (Oracle, Postgres, MySQL...), while others are wrappers to highly-efficient, specific-purpose libraries (bsddb, NetCDF3...). Others have developed their own specific formats to fulfill their own requirements.

In the data-hungry world of scientific computing, one usually prefers (with good reason) solutions that are not only fast but also well-tested and, perhaps more importantly, have outstanding backward and forward format compatibility. Scientific applications also tend to focus on the most efficient ways to find the "needles in the haystack" of massive amounts of data.

We will begin the talk with a description of HDF5 [1], an emerging standard format to store scientific and other data. Its main features will be covered, and the contexts where it can be applied to an advantage will be discussed. We will then introduce PyTables [2], a well-known and widely adopted solution implemented in Python for manipulating potentially huge HDF5 datafiles easily and efficiently.

We also plan to offer a sneak preview of the next-generation PyTables toolkit, with its greatly improved indexing and search capabilities. The PyTables discussion will include benchmarks of the latest versions, to give an idea of its lookup speed and performance as compared to other well-established standard databases and toolkits. Users will learn what they can expect from the next-generation PyTables and how it can help them to find specific data (the needle) in huge (terabytes and petabytes) datasets very rapidly.

In conclusion, we will unfold our master plan for the future domination of the world by PyTables and its growing family.

[1] http://hdf.ncsa.uiuc.edu/HDF5/
[2] http://www.pytables.org/

Primary author: Mr FRANCESC, Altet (Cárabos Coop. V.)

Co-author: Mr IVAN, Vilata (Cárabos Coop. V.)

Presenter: Mr FRANCESC, Altet (Cárabos Coop. V.)

Session Classification: Python in Science

Track Classification: Python in Science