Python tools and services for LHC data analysis

R. De Maria,

C. Hernalsteens, T. Levens, L. Mascetti, D. Piparo

Existing tools

LHC machine data is normally available through:

- the CERN Accelerator Logging Service that can be queried with the Timber application (provided Java 1.8 and jws are installed);
- the NFS file system available from the technical network (e.g. `ssh cs-ccr-dev1`) in the directory `/user/slops/data/LHC_DATA/OP_DATA`;
- in the Post Mortem System database accessible through its java API and other dedicated tools.

Data analysis work flow

Steps and challenges:

- Get data: data in different network, in special databases, deleted after some time
- Explore data: slow plots, difficult to align data, missing points, complex data conditioning
- Save data: when large data sets too many files, too large files, data access becomes to slow, local disks do fail
- Publish results: need finer control on graphics

New tools and services

- PyTimber: Python API for querying data of the CERN Accelerator Logging Service
- EOS ABP space: tree in the EOS filesystem (eos/cern.ch/project/abpdata) which can be accessed from lxplus, lxplus-cb6, web server.
- PageStore: a Python library for storing and retrieving measurement data like, but not limited to PyTimber data.
- Old LHC data collection: stored in EOS and accessible with PageStore.
- SWAN : Jupyter Python notebook server with access to EOS and PyTimber/PageStore installed.

PyTimber

- Developed by R. De Maria, T. Levens, C. Hernalsteens
- Use Java CALS API (BE/CO) through python-jpype and cmmbuild-dep-manager (T. Levens)
- Simple API to get data, search variable, name explore variable tree:

```
import pytimber

db=pytimber.LoggingDB()

data=db.get(“LHC.BOFSU:OFSU_ENERGY”,”2016-05-30 00:00:00.000”,
”2016-05-30 01:00:00.000”)

vars=db.search(“LHC%ENERGY”)

print(db.tree.LHC.Instrumentation)
```

Notebook example

PageStore

- Developed by R. De Maria.
- Store data in indexed binary pages managed by SQLite database:
- Can be queried like PyTimber:

```
import pagestore, pytimber, time
mydb=pagestore.PageStore(‘mydata.db',‘./mydata')
db=pytimber.LoggingDB()
now=time.time()
data=db.get(“%LUMI_TOT_INST”,now-3600*24,now)
mydb.store(data)
mydb.get(“ATLAS:LUMI_TOT_INST”,now-3600,now)
```

EOS ABP Space

- EOS is fast ~100 PB storage space from CERN.
- It has been used as file storage by the experiments in the last few years.
- It is going to replace AFS in the near future as file system.
- It is being the CERNBOX service.
- It can be accessed using eos command line tool, xrootd, as filesystem logging into lxplus-cb6.
- eos command line tool is available in lxplus and as a package in Scientific linux.

EOS team (L. Mascetti et al.) has granted 20TB not back-up space for ABP in:

eos/cern.ch/project/abpdata (soon to be replaced by eos/cern.ch/project/a/abpdata)

Backup will be done on CASTOR: castor/cern.ch/accel/abpdata.

LHC old data repository

In the past many LHC high frequency data was filtered out every week (e.g. BBQ raw data, BLM, Power converter current).

A subset of filtered data from 2010, 2011, 2012, 2015 has been saved for about few TB.

Data is now accessible in EOS through PageStore.

Notebook example

SWAN service

All tools mentioned before can be installed in local machines or in lxplus.

Python code can be developed using command line tools like IPython or web based notebook.

SWAN service has been developed by PH/SFT (D. Piparo et al.) to provide a centrally managed cloud accessible with CERN credential with EOS and Python software installed.

The service is still in beta and few ABP users have granted access, but it will be soon open to everyone.

References and documentation

https://twiki.cern.ch/twiki/bin/view/ABPComputing/LhcDataStorage

accessible from

cern.ch/abp-computing