Boosting your IT infrastructure with SWAN

https://swan.web.cern.ch


CERN

CS3 Workshop
30/01/2017
**Jupyter Notebook**: A web-based interactive computing interface and platform that combines code, equations, text and visualisations.
Access TTree in Python using PyROOT and fill a histogram

Loop over the TTree called "events" in a file located on the web. The tree is accessed with the dot operator. Same holds for the access to the branches: no need to set them up - they are just accessed by name, again with the dot operator.

```python
In [1]:
import ROOT

t = ROOT.TFile.Open("http://indico.cern.ch/event/395198/material/0/0.root");
h = ROOT.TH1F("TracksPt", "Tracks;Pt [GeV/c];#", 128, 0, 64)
for event in t.events:
    for track in event.tracks:
        h.Fill(track.Pt())
c = ROOT.TCanvas()
h.Draw()
c.Draw()
```

Tracks

<table>
<thead>
<tr>
<th>TracksPt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entries</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Std Dev</td>
</tr>
</tbody>
</table>
Jupyter Notebook: A web-based interactive computing interface and platform that combines code, equations, text and visualisations

- ROOT has been fully integrated with the Jupyter technology
  - Data analysis library widely used in HEP (http://root.cern)
- Two flavours: Python and ROOT C++
- JavaScript interactive visualisation
SWAN: Data analysis “as a service”

**Interface:** Jupyter Notebooks

**Goals:**
- Analysis **only with a web browser**
  - Platform independent ROOT-based data analysis
  - Calculations, input and results “in the Cloud”
- **Easy sharing** of scientific results: plots, data, code
  - Storage is crucial: mass & synchronised
- **Simplify teaching** of data processing and programming
- Integration with other **analysis ecosystems**: R, Python, …
SWAN Building Blocks

Storage

Software

Infrastructure
SWAN in the CERN Ecosystem

SWAN federates a set of production technologies at CERN:

- Authentication with **CERN credentials (SSO)**
- Infrastructure: **virtual machines** in OpenStack Cloud
- **Software distribution**: CVMFS
  - Centrally distributed software
- **Storage access**: CERNBox, EOS
  - Experiments’ and users’ data

Plus some external technologies:

- JupyterHub
- Docker
<table>
<thead>
<tr>
<th>Path</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACAT 2016</td>
<td></td>
</tr>
<tr>
<td>CHEP 2016</td>
<td></td>
</tr>
<tr>
<td>cmsdata</td>
<td></td>
</tr>
<tr>
<td>CSC</td>
<td></td>
</tr>
<tr>
<td>ExampleDir</td>
<td></td>
</tr>
<tr>
<td>IMLmeeting</td>
<td></td>
</tr>
<tr>
<td>mylibs</td>
<td></td>
</tr>
<tr>
<td>node_modules</td>
<td></td>
</tr>
<tr>
<td>other</td>
<td></td>
</tr>
<tr>
<td>ROOT-Primer</td>
<td></td>
</tr>
</tbody>
</table>

Same content as in cernbox.cern.ch
Different ways of sharing notebooks

- Share in read-only mode: static visualisation (CERNBox UI)
- Share by link, R/W: concurrent edition not supported by Jupyter
- Notebooks often rely on some input data: share directories
Example notebooks at swan.web.cern.ch

Click on the blue ribbon to open them in SWAN!

Click on the image for a static visualisation
• Strategy to configure the software environment:
  – Docker: *single* thin image, not managed by the user!
  – CVMFS: configurable environment via “views”
  – CERNBox: custom user environment
Offloading to External Resources

In collaboration with CERN Database and Storage groups

User Notebook

Spark Master

Spark Worker

Python task

Python task

Python task

EOS

CernVM File System
• Pilot Service released in June 2016
  https://swan.cern.ch
• Open to everyone with a CERN account
  – And a CERNBox account activated
• Stats of the first 6 months:
  – 4000+ sessions created
  – 8000+ notebooks opened
  – Peak of ~250 sessions
• Service open outside CERN since end of July
Physics Analysis

Rare B meson decay in LHCb

- Read data from EOS
- Setup complex fit
- Document and inspect results

Results coming from real data! (published now)
CERN Beams Department


- Read measurements coming from pick-ups in a database
- Plot time series
- Python, SciPy, Matplotlib
• SWAN as platform for outreach
  – Introductory course about experimental HEP for future high school teachers
• Practical Statistics for Particle Physics Analyses
https://indico.cern.ch/event/545212/

• CERN Summer Student Program: ROOT
https://indico.cern.ch/event/536772/

• CERN School of computing: Parallelization lectures
http://indico.cern.ch/event/502875/

• Data Science @ LHC Workshop, Multivariate analysis tutorial
http://indico.cern.ch/event/395374/
• **SWAN pilot service available**
  – *CERNBox* as a home + *EOS* for mass storage
  – *CVMFS* for software distribution

• **SWAN demonstrated to be a federator of services**
  – Increase added value with respect to individual services

• **Future plans:**
  – Packaging for installation on other sites
  – Improve experience with storage: sharing
  – Enrich CVMFS repositories with new packages (e.g. ML libraries)
  – Exploit external resources (e.g. Spark clusters)
Backup
The W Analysis ROOTbook

The W boson analysis is intended to provide an example for a high statistics analysis using the ATLAS open data dataset. Furthermore, it tests the description of the real data by the simulated W boson data which represents the most extensive dataset in terms of luminosity.

The Z Analysis ROOTbook

Many analyses selecting leptons suffer from Z + jets as a contributing background due to its large production cross section. It is therefore vital to check the correct modelling of this process by the Monte-Carlo simulated data. It is important to measure well-known Standard Model particles, to confirm that we understand properly the detector and software. We are then ready to search for new physics.
Mano S. (14 years old), K12 student

- Approaches programming for the first time
- Verifies numerically what he learned at school
- Shares results with his supervisor and classmates
CERN IT Department

CERN tape system analysis

- Report, data source, analysis and results in a single document