THE 2018 INTERNATIONAL WORKSHOP ON HIGH ENERGY CIRCULAR ELECTRON POSITRON COLLIDER

IHEP Jupyter Service for physics analysis in the future

Tao Lin, Mengyao Qi, Yan Liu, Qiulan Huang, Wei Zheng, Weidong Li
Computing Center, Institute of High Energy Physics, CAS

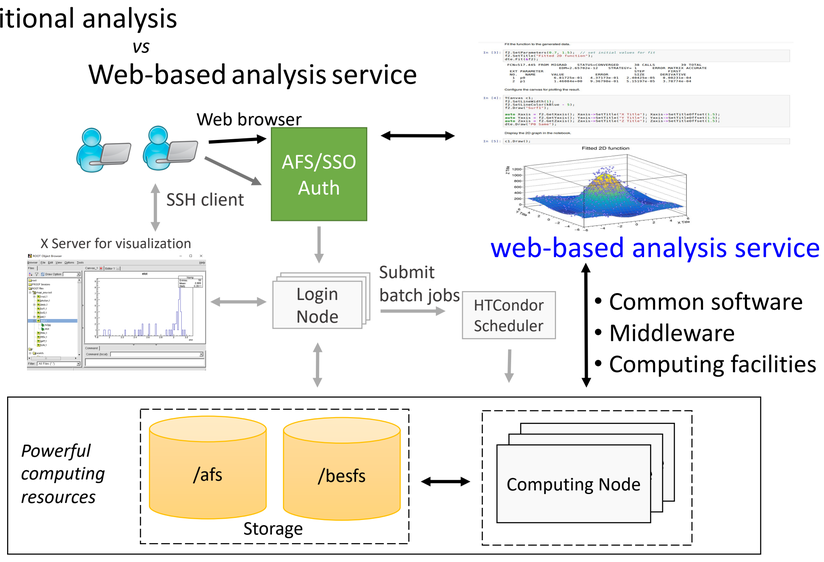
• Motivation

- Physics analysis tasks is challenge due to “big data” (PB scale).
- A physics analysis task is split into many jobs manually, while each job will process a part of dataset.
- For the physics analysis, the same dataset will be loaded many times due to iteration.
  1. Write code, edit job option and update selection criteria.
  2. Submit batch jobs.
  3. After jobs finished, plot data, analyze result.
  4. Go to step 1.
- Submitting jobs and plotting are the most time consuming part.
- Loading dataset many times also cause I/O performance.

Q: How to reduce the iteration time? A: “**Analysis as a Service**”

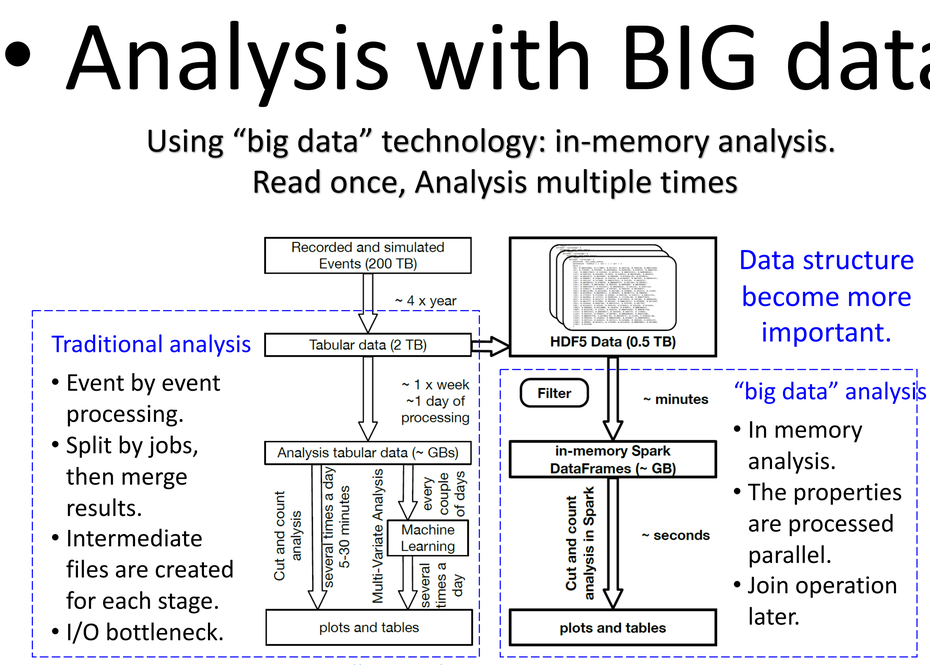
• Analysis workflow

Traditional analysis
vs
Web-based analysis service



• Analysis with BIG data

Using “big data” technology: in-memory analysis.
Read once, Analysis multiple times



Data structure become more important.

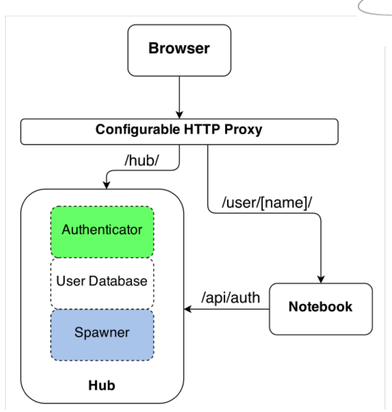
Traditional analysis
- Event by event processing.
- Split by jobs, then merge results.
- Intermediate files are created for each stage.
- I/O bottleneck.

“big data” analysis
- In memory analysis.
- The properties are processed parallel.
- Join operation later.

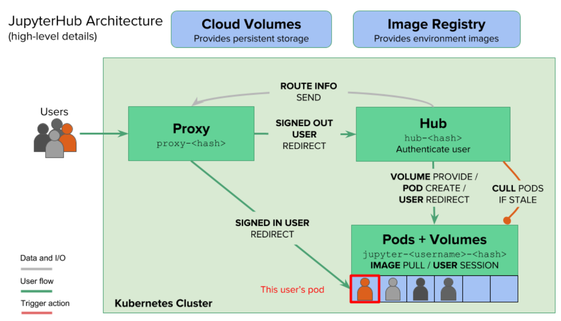
RDataFrame in ROOT: https://root.cern.ch/rdataframe-session-chep-2018-0

• IHEP Jupyter Service

Architecture: jupyterhub + kubernetes

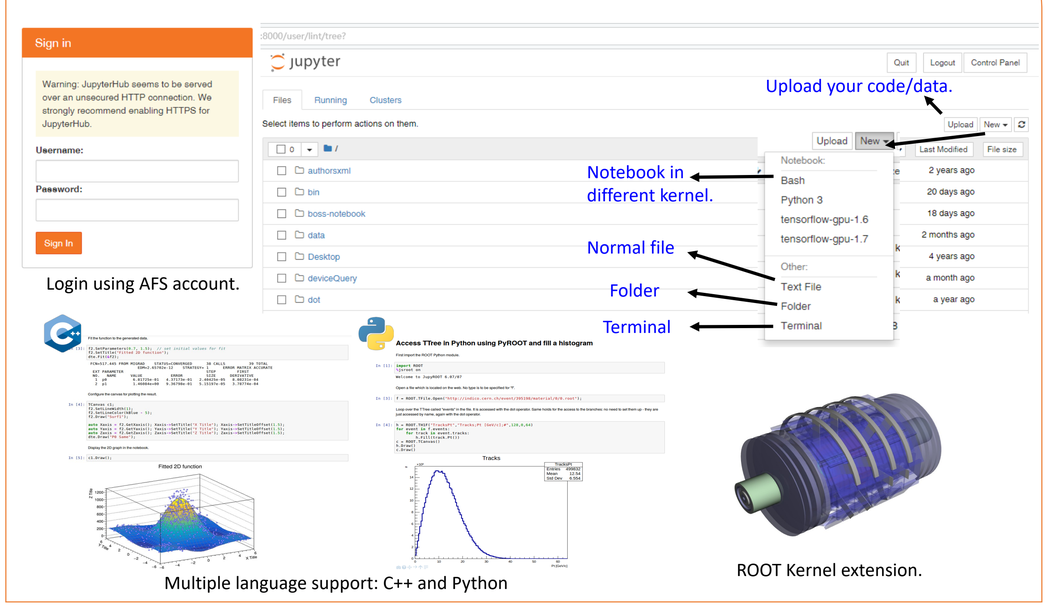


- Kubernetes manages the computing resources.
- On-demand, scalable computing.
- The backend can be also integrated with HTCondor.



- JupyterHub manages multiple instances of the single-user Jupyter notebook.
- Authenticates users.
- Spawns notebooks for users.

• Uses case



Login using AFS account.

Multiple language support: C++ and Python

ROOT Kernel extension.

• Status and Challenges

Status
- Jupyter software stack is deployed at AFS.
- Users can start Jupyter and ROOT 6 in their own space.
- Setup JupyterHub in a virtual machine and setup Kubernetes in two blade servers.

Challenges
- Need to klog manually to get AFS token.
- Support multiple users and experiments.
- Unified data access between SSH and Web.
- How to benefit from “big data” technology.

Make the service available as soon as possible!
Get more feedbacks from physicists.

References:

1. SWAN: http://dx.doi.org/10.1016/j.future.2016.11.035
2. DIANA: http://diana-hep.org
3. Jupyter: http://jupyter.org/
4. JupyterHub: https://jupyterhub.readthedocs.io/en/latest/index.html
5. Jupyter and Spark: https://mapr.com/blog/configure-jupyter-spark-python/
6. Fermilab big data: http://computing.fnal.gov/big-data/
7. NERSC: http://www.nersc.gov/users/data-analytics/data-analytics-2/jupyter-and-rstudio/
8. Spark for HPC: https://ieeexplore.ieee.org/document/7965154/
9. ROOT RDataFrame: https://indico.cern.ch/event/743070/

• Conclusions

- Jupyter service will be available at IHEP, providing users web based interactive analysis.
  - A general solution to support different users and experiments.
  - It can speedup the analysis such as in-memory big data analysis.
- A prototype based on JupyterHub is deployed in a virtual machine.
  - User authentication works, however AFS token does not work properly.
- Kubernetes cluster is also setup using two blade servers.

Authors: Tao Lin, Mengyao Qi, Yan Liu, Qiulan Huang, Wei Zheng, Weidong Li
Computing Center, Institute of High Energy Physics, CAS
Emails: lintao@ihep.ac.cn, qmy@ihep.ac.cn