

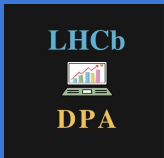
Analysis preservation in HEP

It matters and we can do better

(with a heavy LHC(b) bias)

UK-HEP 2024

N. Skidmore
Nov 2024



Story time...



Me: 2014 - 2017

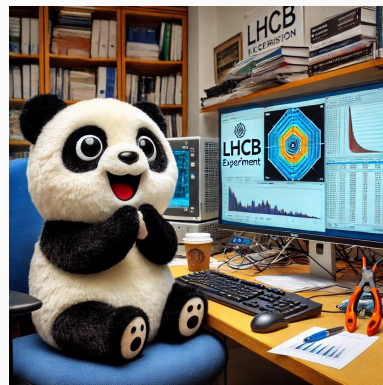
Why do we need to preserve?

Analyses now take longer than a PhD (> 4 years)

- The analysis will have to be “handed over” to a new student (or orphaned 😞)

More complex analyses require more people

- High person-turnover is a feature of our fields
- Analysis preservation infrastructure aids collaboration



What do we need to preserve?

The physics

Analysis note

The data

nTuples

Data provenance

The code and how to run it

Git repo **Documentation**

Workflow manager

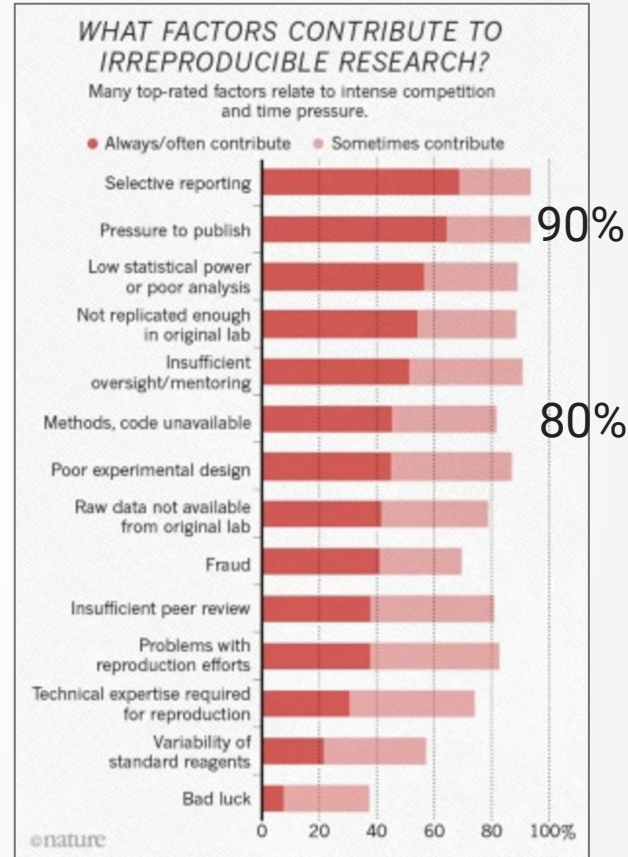
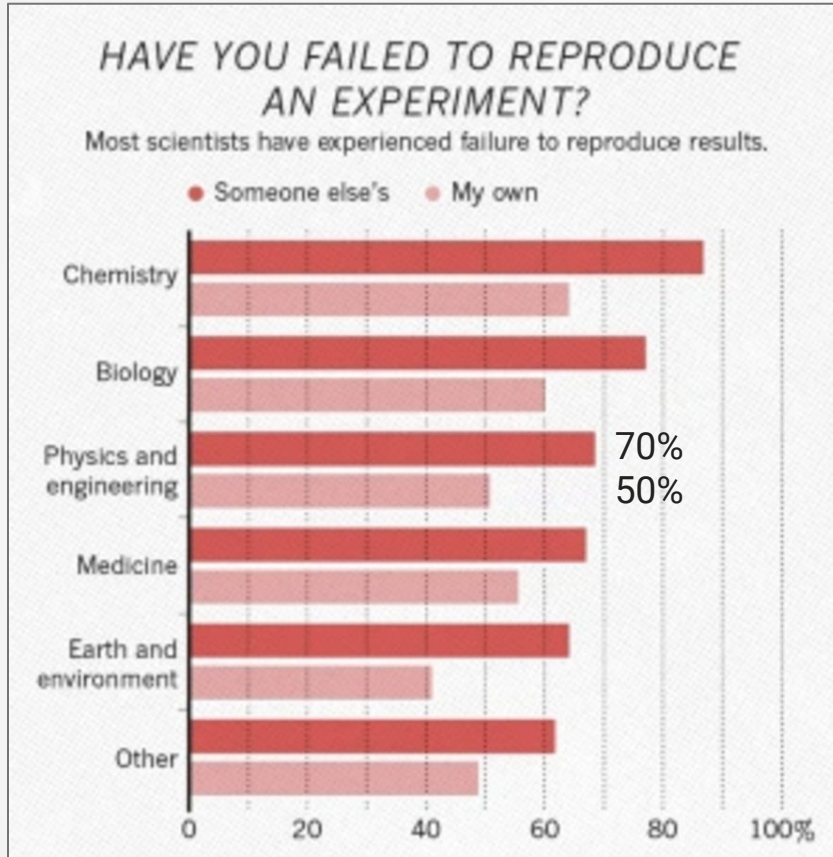
The ability to run the code

GitLab CI

Env preservation

We are bad at sharing code

www.nature.com/articles/533452a



We are bad at sharing code

Theorems, lemmas, corollaries...

We are bad at sharing code

“””The proof is too ugly to show anyone else. It would be too much work to rewrite it neatly so that others could read it.”””

We are bad at sharing code

“””I didn't actually prove the theorem - my student did. They have since graduated and now work as a Quant. But the student was very good... I'm sure the proof was correct.”””

We are bad at sharing code

“””Giving the proof to my competitors would be unfair to me. It took years to prove this theorem, and the same idea can be used to prove other theorems.”””

We are bad at sharing code

Table 10: Top Reasons Not to Share Code

| | Not Share |
|--|-----------|
| → <i>The time it takes to clean up and document for release</i> | 77.78% |
| <i>Dealing with questions from users about the code</i> | 51.85% |
| <i>The possibility that your code may be used without citation</i> | 44.78% |
| The possibility of patents or other IP constraints | 40.00% |
| Legal barriers, such as copyright | 33.72% |
| <i>Competitors may get an advantage</i> | 31.85% |
| → <i>The potential loss of future publications using this code</i> | 31.11% |
| The code might be used in commercial applications | 28.15% |
| Availability of other code that might substitute for your own | 21.64% |
| → <i>Whether you put in a large amount of work building the code</i> | 20.00% |
| Technical limitations, ie. webspace platform space constraints | 20.00% |

+ **Student left**

+ **Self-conscious about code**

10 rules for Reproducible Computational Research

Rule 1: For Every Result, Keep Track of How It Was Produced

Rule 2: Avoid Manual Data Manipulation Steps

Rule 3: Archive the Exact Versions of All External Programs Used

Rule 4: Version Control All Custom Scripts

Rule 5: Record All Intermediate Results, When Possible in Standardized Formats

Rule 6: Always Store Raw Data behind Plots

Rule 7: Generate Hierarchical Analysis Output

Rule 8: For Analyses Including Randomness, Record Random Seeds **Important for pheno!**

Rule 9: Connect Textual Statements to Underlying Results

Rule 10: Provide Public Access to Scripts, Runs, and Results

Unrealistic
with big
datasets

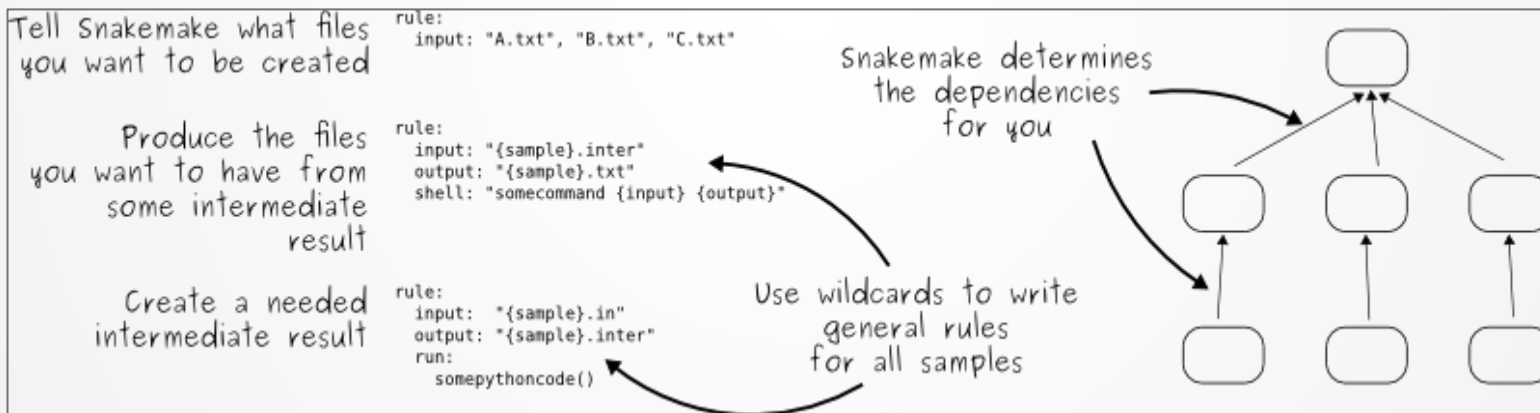
Rule 1: For Every Result, Keep Track of How It Was Produced

Workflow managers

- Breaks analysis into bite-size rules with input, output and command
- Preserves:
 - How to run scripts
 - How every intermediate result is produced (the workflow)
 - The dependency between analysis stages



Snakemake



Rule 3 - Archive the Exact Versions of All External Programs Used

Containers and virtual environments

- Containers encapsulate a computing environment including OS
- Virtual environments only encapsulate Python dependencies

Questions for a new collaborator

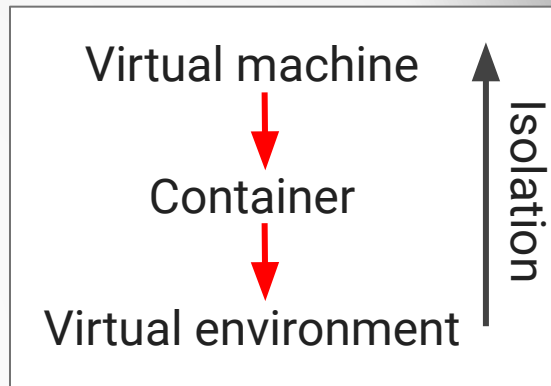
1. How do I login?
2. How do I access the data?
3. How do I setup my environment?
4. How do I get the analysis s/w libraries?
5. How do I run the analysis steps?

Containers



Virtual

environments



- These do take some time and setup but they enable
- Quick onboarding
 - Efficient collaboration

Rule 3 - Archive the Exact Versions of All External Programs Used

Containers and virtual environments

- Containers encapsulate a computing environment including OS
- Virtual environments only encapsulate Python dependencies

Virtual machine

Container

Isolation

“Well it works on my machine...”

Questions

1. How do I login?
2. How do I access the data?
3. How do I setup my environment?
4. How do I get the analysis s/w libraries?
5. How do I run the analysis steps?

Containers

**Virtual
environments**

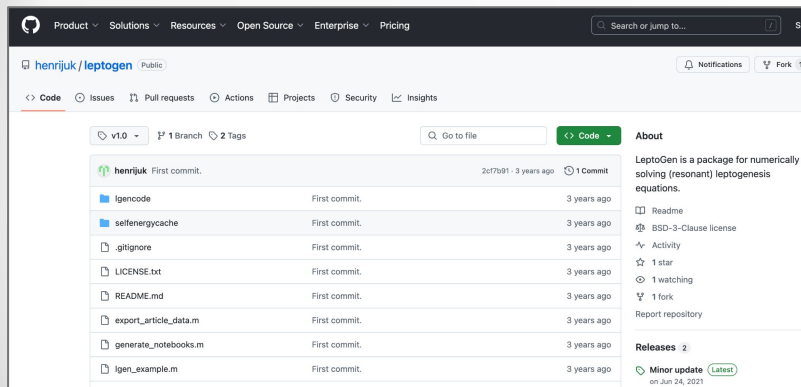
These do take some time and setup but they enable

- Quick onboarding
- Collaboration
- Publishable software
- Reuse and repurposing

Rule 3 - Archive the Exact Versions of All External Programs Used

Package and version software

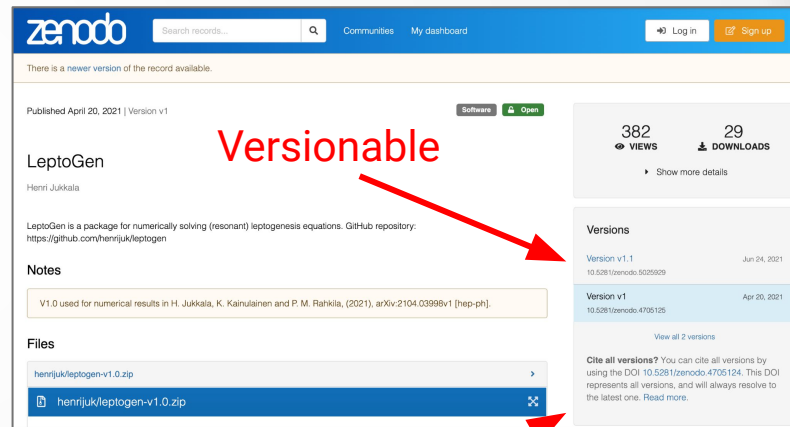
- Encourage packaging, versioning and long term maintenance of software tools



Archive
the tag of
a repo



zenodo.org/records/4705125



Versionable

Persist identifiers - DOIs

Note added

Arxiv paper

The Mathematica code package that was used to compute all numerical results in this paper is publicly available at <https://doi.org/10.5281/zenodo.5025929>.

Rule 3 - Archive the Exact Versions of All External Programs Used

Package and version software

pypi.org/project/triggercalib/

PyPI package
installable with pip

triggercalib 1.3.5

✓ Latest version

Released: Sep 24, 2024

`pip install triggercalib`

Tooling for data-driven efficiencies of the LHCb trigger

Navigation

- Project description
- Release history
- Download files

Verified details

These details have been verified by PyPI

Maintainers

JamieGooding

Unverified details

These details have not been verified by PyPI

Project links

- Homepage
- Bug Tracker

Project description

TriggerCalib - Tooling for trigger efficiencies

This repository contains tools developed for calculating trigger efficiencies in LHCb analyses and studies. The full documentation for the TriggerCalib tools can be found here: <https://triggercalib.docs.cern.ch/>

At the core of these tools is the `HTEff` class, which implements the TISTOS method (as laid out in [LHCb-PUB-2014-039](https://arxiv.org/abs/2014.039)) to produce trigger efficiencies in ROOT TH1/TH2 histograms. This will be extended in the near future by a `yaml`-configurable interface to the class, with the aim of being familiar to users of the `HTEfficiencyChecker` tool for studying MC efficiencies in simulation. An additional tool, currently in plan, will further extend this functionality by providing users with trigger efficiency correction tables (à la `PEPCal1b2`) for control channels.

If you wish to contribute to TriggerCalib, please see [CONTRIBUTING.md](https://github.com/triggercalib/triggercalib/blob/main/CONTRIBUTING.md).

Acknowledgements

We acknowledge funding from the European Union Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

SMARTHEP

EUROPEAN UNION

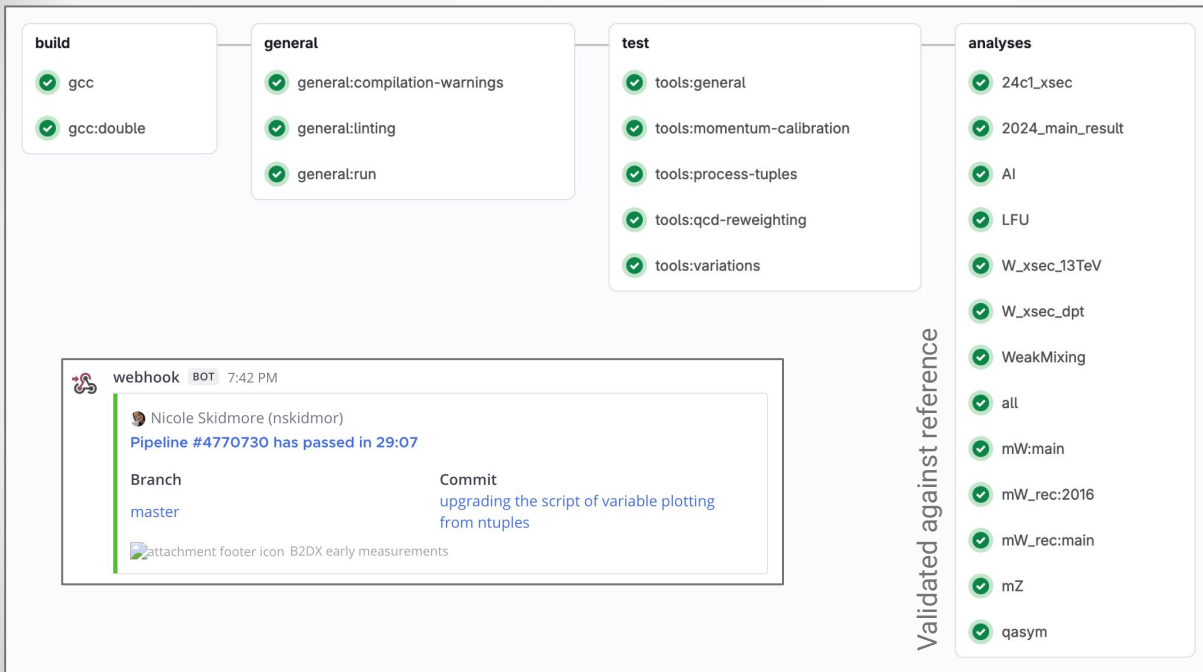
MARIE CURIE ACTIONS

Encourage
contributions

Can acknowledge
funding agencies

Rule 4: Version Control All Custom Scripts

Git repo and CI



WeakMixing-run-2-unblinding protected
12dc96f8 · Merge branch 'mvesteri-weak-mixing-unblind' into 'master' · 5 months ago

Project at the unblinding of the WeakMixing analysis (24th July 2024)

mW-2016-unblinding protected
e48b2866 · Merge branch 'mpilL_mW_summary_plot' into 'master' · 3 years ago

Version of the code for the unblinding of the W mass measurement using 2016 data on 08/06/2021

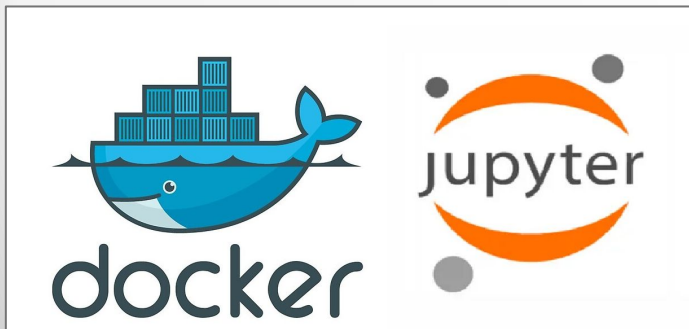
ana-note-v1.0
e1281d74 · Merge branch 'mramospe-qcdbg-d-note' into 'master' · 3 years ago

version of the ANA note sent to QEE conveners

Rule 9: Connect Textual Statements to Underlying Results

Notebooks

- Notebooks provide a way to connect physics reasoning and code
- Can run notebooks in containerised analysis environments



Notebook example with MC

Lets look at some $B \rightarrow D\pi$ MC data using RDataFrame

- Plot the B and D mass

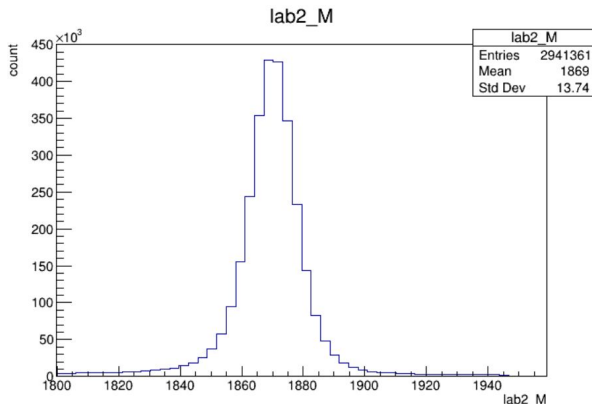
```
In [8]: import ROOT
ROOT.gErrorIgnoreLevel = ROOT.kInfo

files = ["/eos/lhcb/wg/b2oc/TD_Bs2Dsh_Run2/MC_Run2/Stripping34_Sim09h/B2DX_MC_11264001_Bd_D-pi_2018_dw.root"]

ntupleName = "Bd2DPiOfflineTree/DecayTree"
fileName = files[0]
dataframe = ROOT.RDataFrame(ntupleName, fileName)

canvas = ROOT.TCanvas()
Dmass=dataframe.Filter("lab0_M>4800 & lab0_M<6000").Histo1D("lab2_M")
Dmass.GetXaxis().SetRangeUser(1800, 2000)
Dmass.Draw()
canvas.Draw()
```

SWAN example



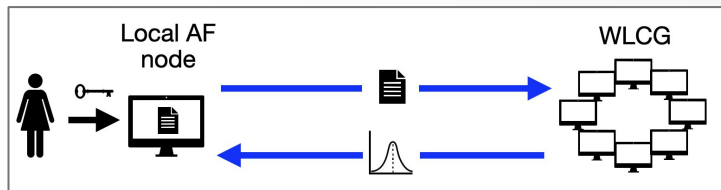
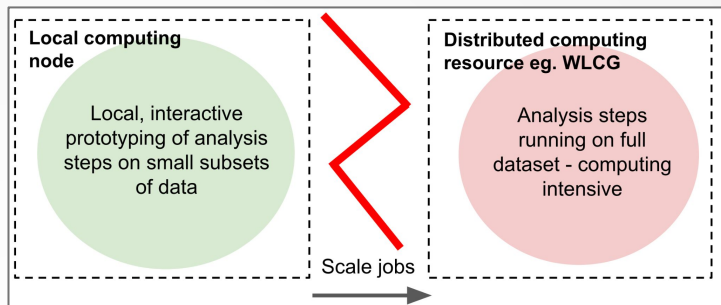
Rule 9: Connect Textual Statements to Underlying Results

Analysis facilities

"Infrastructure providing the data, software and computational resources to execute (an element of) an analysis workflow. ... shared and supported through virtual organization."

154th LHCC Meeting

"The LHCC recommends that experiments engage in the process of developing and defining the structure of the future Analysis Facilities"



- AFs allow automatic, transparent **scaling** to batch resources from interactive session (notebook)
- Authentication, submission and retrieval abstracted away from user
- Results returned as if job was run locally

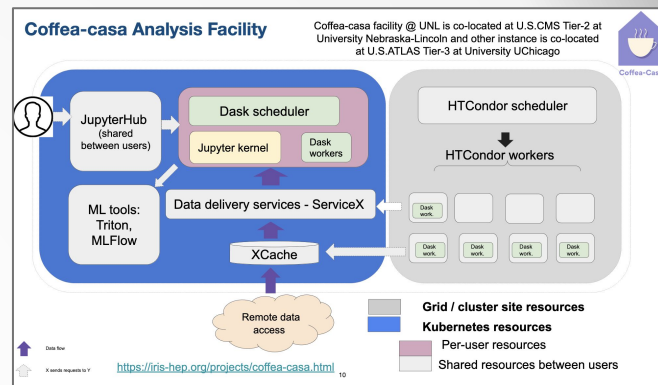
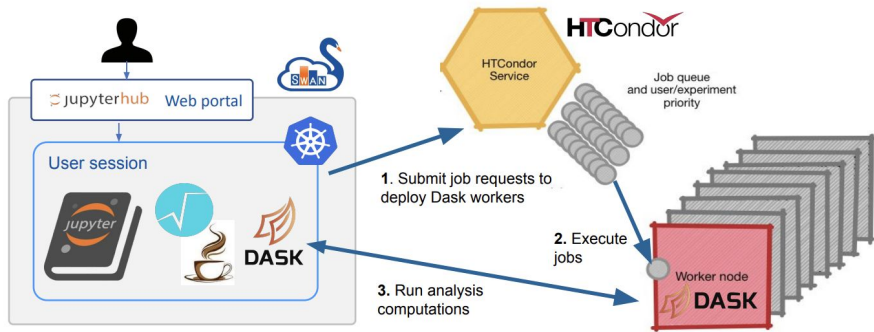
Rule 9: Connect Textual Statements to Underlying Results

AF infrastructure for ATLAS and CMS US analysts is advanced

- Jupyterhub (interactive notebooks)
- Integrated Dask scheduler for scaling to batch resources
- Token based AAI

Focus on **scale out of interactive analysis**

- On **already existing** CERN Batch system resources
- Via RDataFrame / coffea + Dask



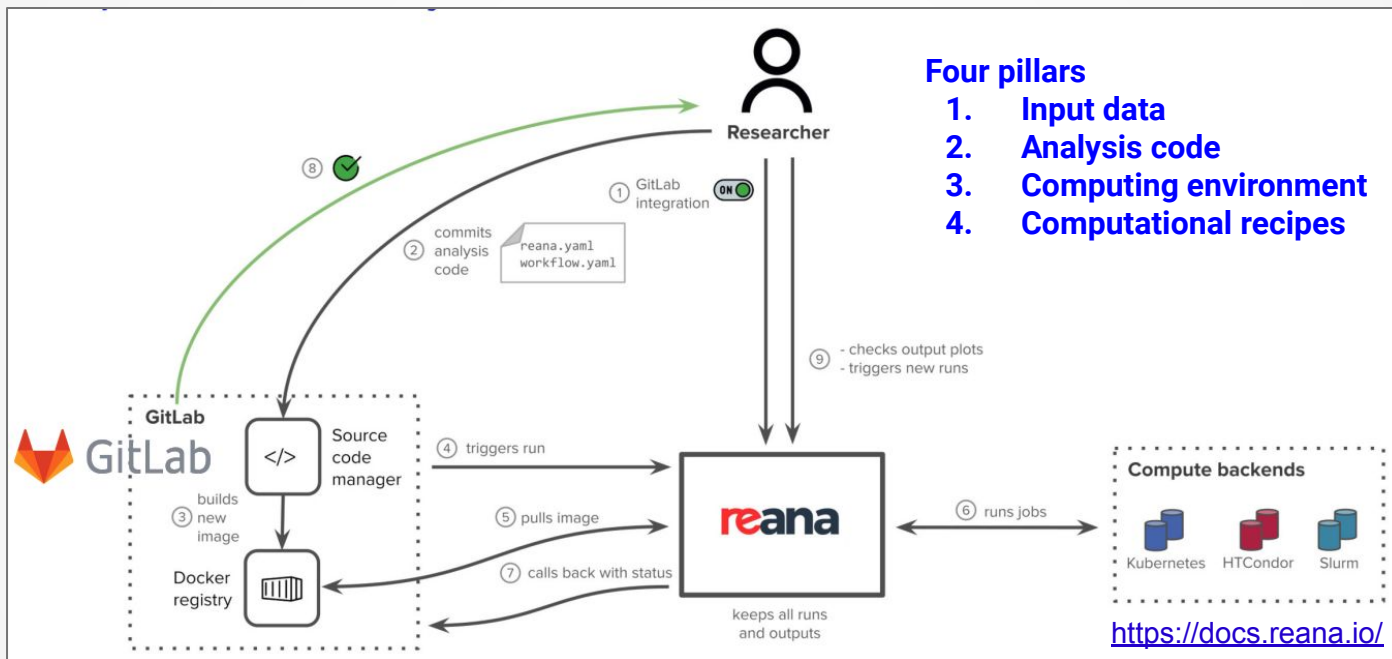
CERN's SWAN service has deployed **AF pilot**

"Interactive computing on big datasets, with analysis built on frameworks like RDataFrame and coffea"

SWAN AF talk at CHEP

Rule 10: Provide Public Access to Scripts, Runs, and Results

REANA - REproducible ANALysis platform that can run containerised analysis pipelines on remote computing resources



AI to the rescue?



AI to the rescue?

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.integrate import quad
```

```
def sawtooth_wave(x):
    return x / 3.14
```

```
x=[]
for i in range(100):
    x.append(-3.14+i/1000.)
x=np.array(x)
plt.plot(x, sawtooth_wave(x))
plt.show()
```

✓ 0.5s

AI to the rescue?

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.integrate import quad
```

```
def sawtooth_wave(x):
    return x / 3.14
```

```
x=[]
for i in range(100):
    x.append(-3.14+i/1000.)
x=np.array(x)
plt.plot(x, sawtooth_wave(x))
plt.show()
```

✓ 0.5s

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.integrate import quad
```

```
def sawtooth_wave(x):
    """
    Generates a sawtooth wave function.

    Args:
        x: The input x-values.

    Returns:
        The sawtooth wave function.
    """
    return x / np.pi
```

```
# Define x-axis values
x = np.linspace(-np.pi, np.pi, 1000)
```

```
# Plot the sawtooth wave
plt.plot(x, sawtooth_wave(x))
plt.xlabel("x")
plt.ylabel("y")
plt.title("Sawtooth Wave")
plt.grid(True)
plt.show()
```

pi not
hardcoded

Doc-string with
purpose and
arguments

Use of numpy
arrays allowing
vectorisation

Plot labels and
formatting

AI to the rescue?

But not with the physics

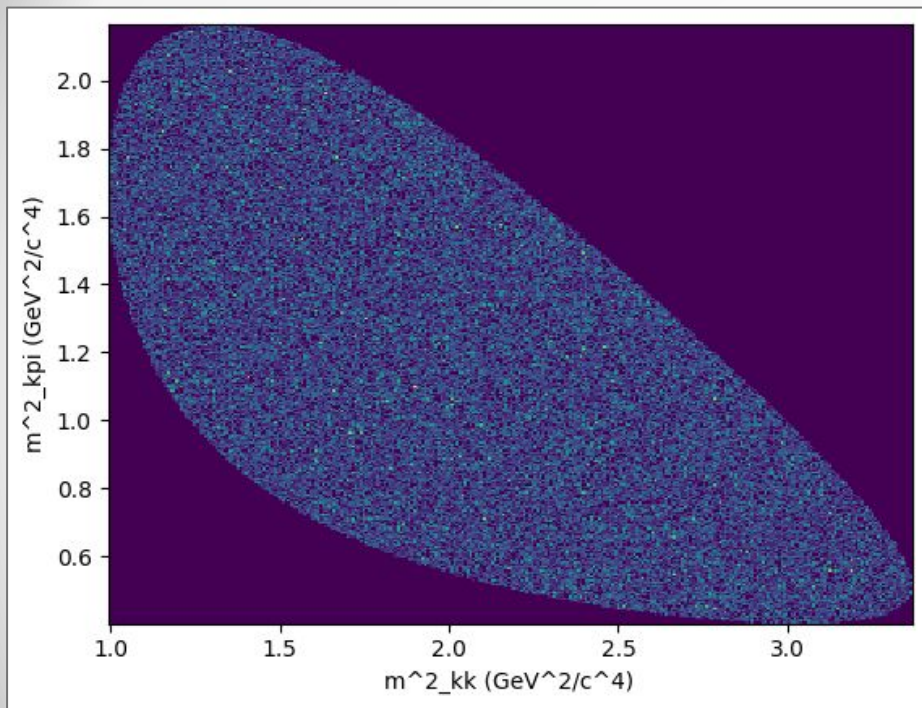
Task 1

Write a function(s) that generates n decay events of $D_s^+ \rightarrow K^+ K^- \pi^+$ that are uniformly distributed over the allowed phasespace. The parameter n should be an argument of your function. Plot these events on a Dalitz plot for $n = 100000$ using a `hist2d` with 300x300 bins.

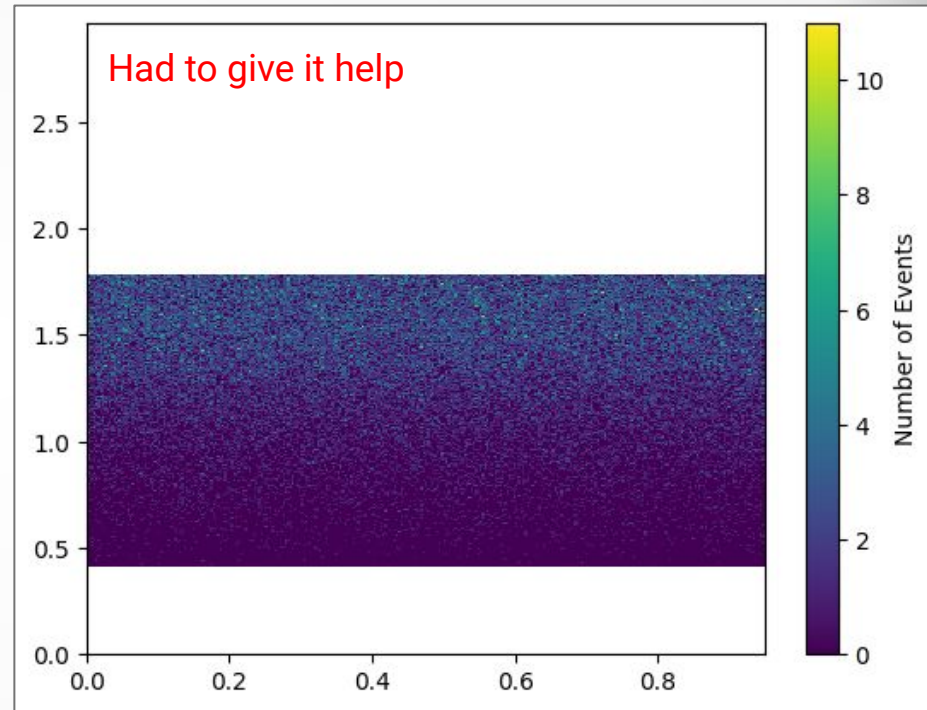
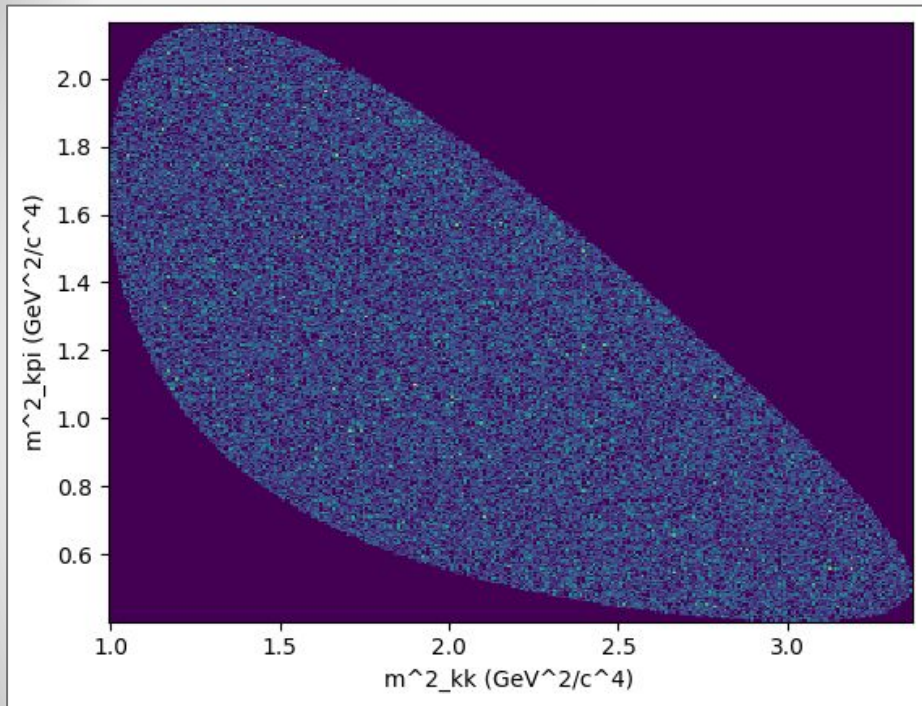
- "Allowed phasespace" here refers to the kinematically allowed region of the Dalitz plot, ie. the area shown in the example Dalitz plots above.
- Use the same units (GeV^2/c^4) and axes as in the example Dalitz plot on the left. To ease communication and make the code clearer we will use $d \rightarrow abc$ notation for the decay such that you will be working with the variables:
 - `m2ab` - the invariant mass squared of the $K^+ K^-$ system
 - `m2bc` - the invariant mass squared of the $K^- \pi^+$ system
 - `md, ma, mb, mc` - the masses of the decaying D_s meson and the 3 decay products respectively
- The masses you should use are
 - `m_dmeson=1.97`
 - `m_kaon=0.498`
 - `m_pion=0.135`

Hint: Use the accept-reject Monte Carlo method with NumPy functions. You only need to generate `m2ab` and `m2bc` values for the events. The range of `m2bc` values kinematically allowed given a `m2ab` value can be found in eqn 49.23a and 49.23b of the [PDG review](#) (note that the PDG uses $M^+ \rightarrow 1^+ 2^+ 3^-$ notation)

AI to the rescue?



AI to the rescue?

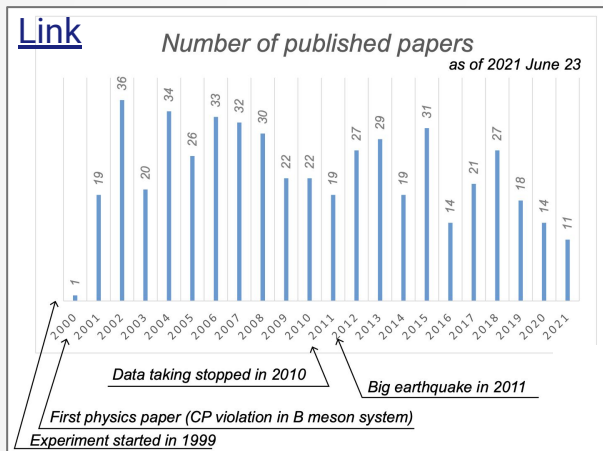


The long term...

The datasets we use are too valuable to not enable their full exploitation by **all**

- Some are completely unique and may remain so for some time eg. SMOG2
- Internal work on **preserving and opening** these must happen now

Belle I papers



Last month...

Search for $h_b(2P) \rightarrow \gamma \chi_{bJ}(1P)$ at $\sqrt{s} = 10.860$ GeV

Belle Collaboration • A. Boschetti [Show All\(142\)](#)

Oct 21, 2024

e-Print: [2410.16181](#) [hep-ex]

Report number: Belle Preprint 2024-07; KEK Preprint 2024-19

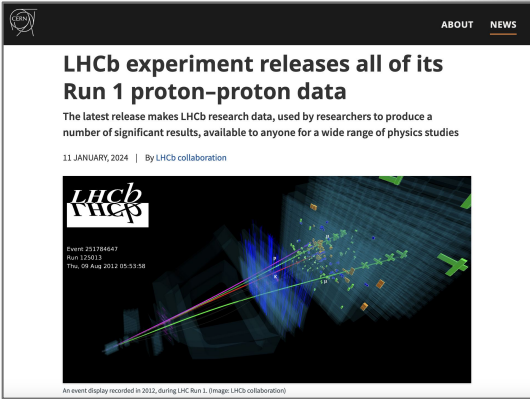
Experiments: [KEK-BF-BELLE](#)

View in: [ADS Abstract Service](#)

Making data open

The [CERN Open Data Policy](#) encourages the release of reconstructed data

- Data from all LHC experiments released through the Open Data Portal
- LHCb released its full Run 1 dataset ~ 800TB



[Link](#)

Releases for Run 2 and beyond impossible due to data volume - not scalable!

| | ALICE | ATLAS | CMS | LHCb |
|-------|-------|--------|------|----------------------------|
| Run-2 | 2 PB | 0.5 PB | 2 PB | 10 PB (including Run-1) |
| Run-3 | 4 PB | 1 PB | 4 PB | 45 PB |
| Total | 6 PB | 1.5 PB | 6 PB | 55 PB |

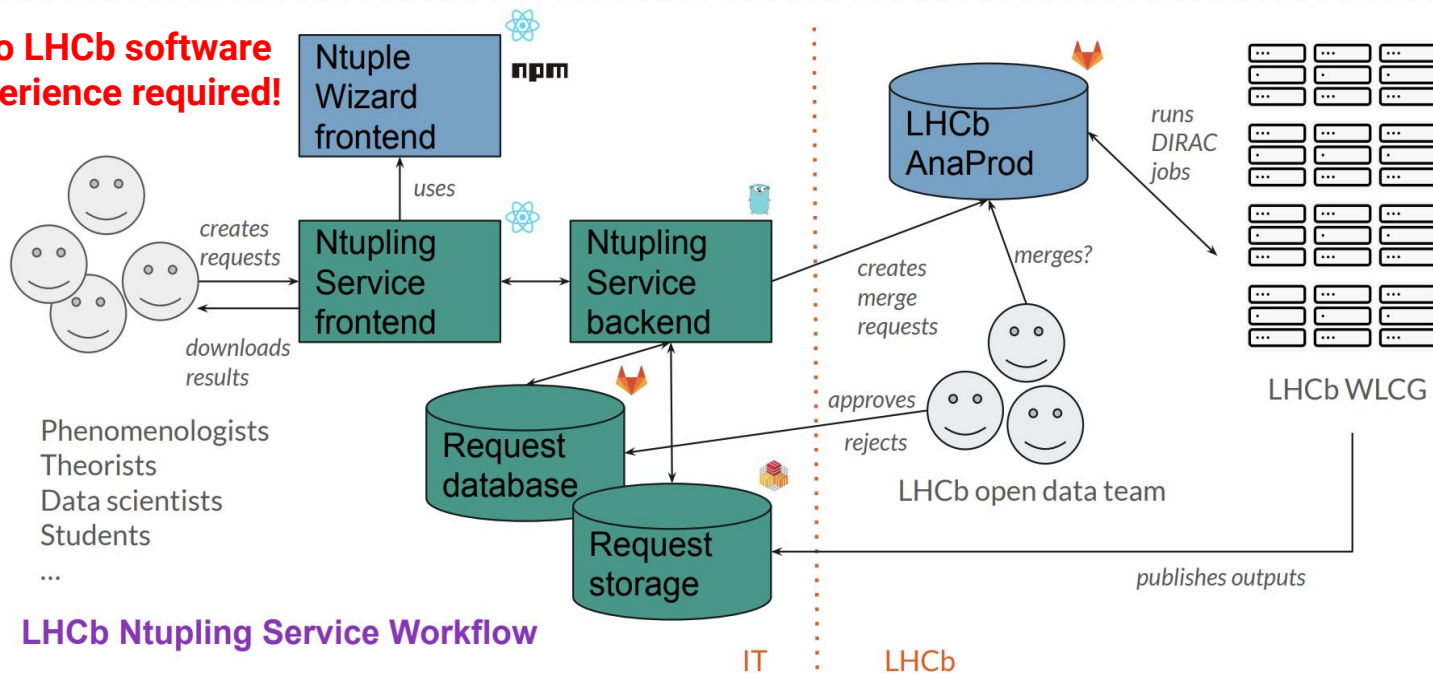
Making data open

Scalable solution - NTuple Wizard!

[CHEP talk](#) on NTuple Wizard

- Developed in close collaboration of the CERN IT department and the LHCb Collaboration!

Zero LHCb software experience required!



Making data open

Scalable solution - NTuple Wizard!

CHEP talk on NTuple Wizard

- Developed in **close collaboration** of the **CERN IT department** and the **LHCb Collaboration!**

Zero
exper

Be part of the beta release! (Run 1 data)

<https://opendata-lhcb-ntupling-service.app.cern.ch/>

<https://indico.cern.ch/event/1429526/timetable/>

Theorists
Data scientists
Students
...

database

Request
storage

LHCb open data team

publishes outputs

LHCb Ntupling Service Workflow

IT

LHCb

Tutorials

hsf-training.org/training-center/

- Git
- CI/CD
- Docker
- Singularity/Apptainer
- Reana
- Analysis essentials incl. Snakemake
- Julia

Self-study and in-person events

| | | |
|---|---|--|
|  | Advanced git Learn to work with branches and more with this interactive webpage. |  GitHub |
|  | CI/CD (gitlab) Continuous integration and deployment with gitlab: automatically run unit tests and more for every commit that you push on gitlab. |  GitHub  Videos |
|  | CI/CD (github) Continuous integration and deployment with GitHub actions: automatically run unit tests and more for every commit that you push on GitHub. |  GitHub  Videos |
|  | Docker Introduction to the docker container image system. Docker allows to consistently run your code in any environment or on any machine, making it an important ingredient to analysis preservation. |  GitHub  Videos |

Close to home...

“Your closest collaborator is you six months ago...
but you don’t reply to email.”

Karl Broman

“Tools for Reproducible Research”

Backup

Analysis productions - declarative ntupling

Pre - run 3 analysts made their own ntuples

THE PROBLEM

- Submitting, monitoring and error handling $O(10,000)$ grid jobs
- No data provenance
- Thousands of failing grid jobs

⇒ BIG barrier between analysts and data



Analysis productions - declarative ntupling

THE SOLUTION ⇒ Analysis productions

- Centralise and automate ntuple creation
⇒ Saves countless analyst-hours
- Exploit DIRAC transformation system
⇒ Full data provenance
- Full job testing on GitLab CI
⇒ No buggy jobs on grid

Simple yaml job configuration

```
defaults:
  application: DaVinci/v64r10@x86_64_v2-el9-clang16-opt
  output: DATA.R00T
  options:
    entrypoint: bs2dspirun3.dv_simple:alg_config
    extra_options:
      input_raw_format: 0.5
      input_type: R00T
      simulation: False
      data_type: "Upgrade"
      geometry_version: run3/trunk
      conditions_version: master
      input_process: "TurboPass"
      input_stream: "b2oc"
  inform:
    -
  wg: B20C

{% set datasets = [
  ('2024Data', 'MagDown', '24c2'),
  ('2024Data', 'MagUp', '24c2'),
```

What application
to run

Job options

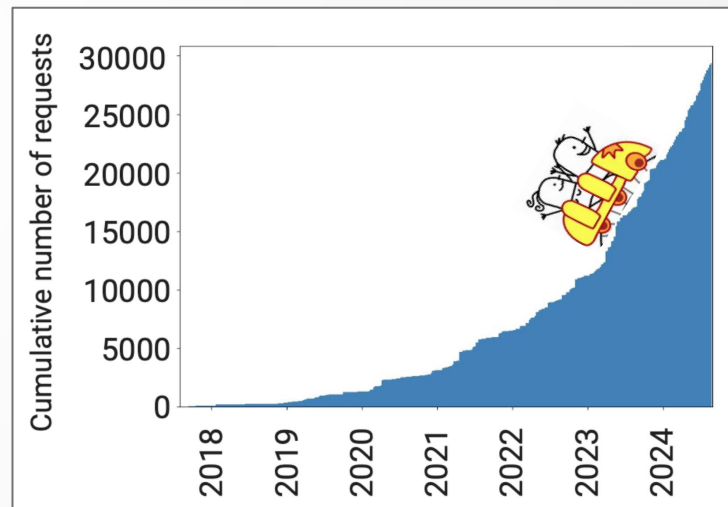
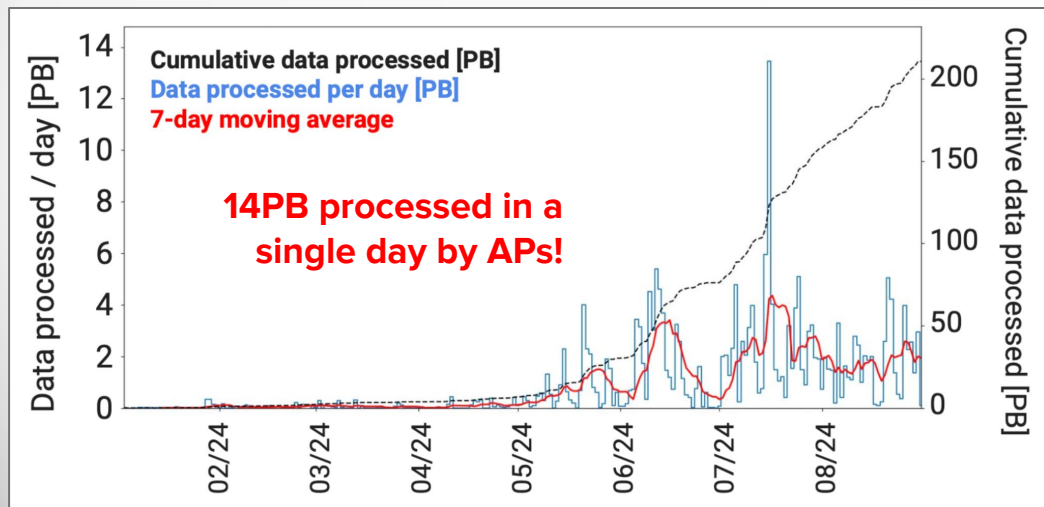
Data to run on

Use Jinja templating to “render” the YAML

Analysis productions - declarative ntupling

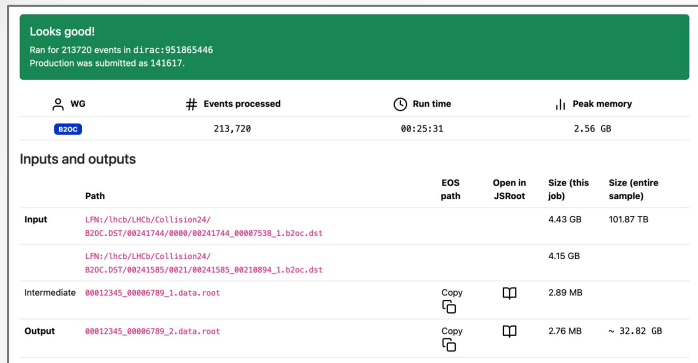
Full adoption of analysis productions at LHCb

- Over 1200 Run 3 APs have been submitted so far
- 700+ “live” APs picking up data as it was Spruced
 - Analysts have been looking at data tuples days after it was recorded by detector
- We are making amazing use of the WLCG!

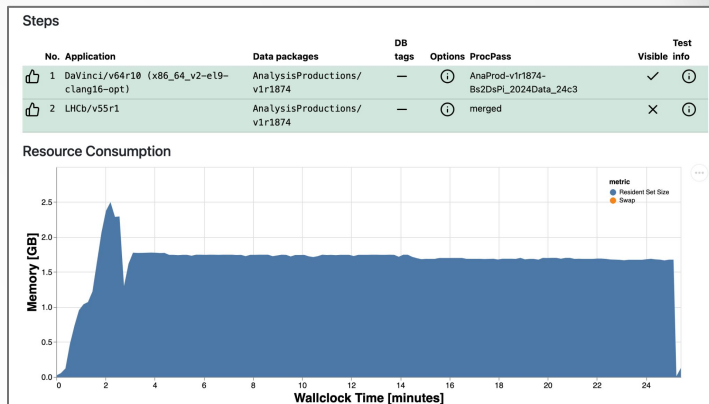


Analysis productions - declarative ntupling

Comprehensive job testing through GitLab pipelines



Reporting on memory usage



Reporting on estimated output size

LHCb Data Processing and Analysis @lhcbdp - 1 week ago Maintainer

Welcome to Analysis Productions!

This is a summary of the productions requested in this merge request:

| Step | Production ID | Num Test LFNs | Run time | Estimated Output Size (MB) |
|------------------------------|---------------|---------------|----------------|----------------------------|
| Bs2DsPL2024Data_MagDown_24c2 | 123713 | 2 | 0:42:06.715033 | 3.1 |
| Bs2DsPL2024Data_MagUp_24c2 | 123714 | 2 | 0:40:23.311033 | 3.0 |
| Bs2DsPL2024Data_MagUp_24c3 | 123715 | 2 | 0:25:31.108833 | 2.6 |
| Bs2DsPL2024Data_MagDown_24c3 | 123716 | 2 | 0:19:40.963231 | 2.5 |
| Bs2DsPL2024Data_MagDown_24c4 | 123717 | 2 | 0:35:30.184766 | 3.2 |

Interactive logs with warning/error highlighting

Logs show less output

DaVinci_1.log prmon_1.txt prodConf_DaVinci_1.json LHCb_2.log prmon_2.txt prodConf_LHCb_2.json DIRAC.log

Copy Download

```
1 Overriding DIRACSYS_CONFIG to /tmp/tmp1_mfqf08/tmp/pilot.cfg
2 Restarting process with ['/cvmfs/lhcb.cern.ch/lhcbdirac/versions/v11.0.48-1727212764/Linux-x86_64/bin/dirac-production-request-run-local', '/t
3 Executing workflow locally
4 Executing from /tmp/951865446
5 Executing job at temp directory /tmp/951865446/Local_99hwjh07_JobDir
6 File not found Request_0_AnalysisProduction_AnaProd-v1r1874-Bs2DsPL2024Data_24c3_EventType_94000000_B20C_1.xml
7 Job has input data requirement, will attempt to resolve data for DIRAC.LocalProdTest.local
8 Replica Lookup Time: 0.48 seconds
9 Metadata Lookup Time: 0.12 seconds
10 Job has a specific policy setting: DIRAC.WorkloadManagementSystem.Client.DownloadInputData
```

Analysis productions - declarative ntupling

Full data
provenance with
datasets tagged
by analysis

Tree display

This section displays the samples split by tags and is the recommended way of requesting datasets. Clicking on one of the boxes will filter the list of samples shown below. See TODO for more information.

Grouped tags ☒ config ☒ datatype ☒ eventtype ☒ polarity

Drag to sort config datatype eventtype polarity

| hs_ft_run2 60 | | | | hsb 16 | | | |
|---------------|--|------------|--|------------|--|------------|--|
| mc 44 | | 13264031 4 | | 13264021 4 | | 13264031 4 | |
| 2016 12 | | 13264021 4 | | 13264021 4 | | 2015 4 | |
| magdown 2 | | magup 2 | | magdown 2 | | magdown 2 | |
| 13164042 4 | | magup 2 | | 13164042 4 | | magup 2 | |
| magdown 2 | | magup 2 | | magdown 2 | | magup 2 | |
| 2017 12 | | 13264031 4 | | 2015 8 | | 2017 4 | |
| 13264021 4 | | 13264021 4 | | 13264021 4 | | 90000000 4 | |
| magdown 2 | | magdown 2 | | magdown 2 | | magdown 2 | |
| 13164042 4 | | magup 2 | | 13164042 4 | | magup 2 | |
| magdown 2 | | magup 2 | | magdown 2 | | magup 2 | |
| 2018 4 | | 90000000 4 | | 2018 4 | | 90000000 4 | |
| magdown 2 | | magdown 2 | | magdown 2 | | magdown 2 | |



apd python packages allows
for easy data file retrieval.
Snakemake integrations!

```
1 from apd import AnalysisData
2
3 datasets = AnalysisData("b2oc", "bs2dsp_i_run3")
4 bs2dsp_i_2024data_magdown_24c2_pfns = datasets(polarity="magdown", eventtype="94000000", datatype="2024")
```