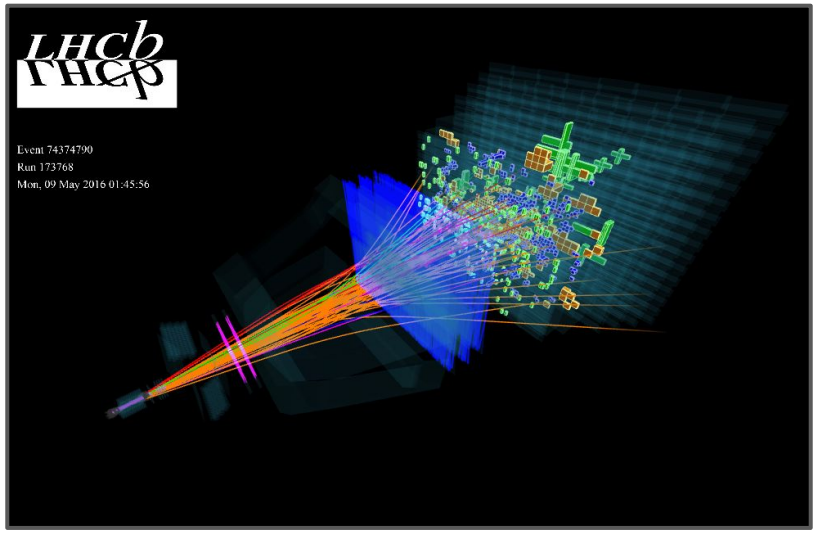


LHCb Analysis Preservation and Open Data Activities

03.10.2024

Dillon Fitzgerald on behalf of the LHCb Collaboration

LHCb Data Processing & Analysis (DPA) Project

The LHCb Data Processing & Analysis Project includes work packages (WPs) related to data preservation, analysis preservation, and open data

- WP2: Analysis Productions
- WP6: Analysis Preservation & Open Data

- WP6 - Analysis Preservation & Open Data
  - Analysis preservation guidelines
  - Analysis workflow preservation - best practices
  - Open data release
  - WG feedback
  - Ntuple Wizard
  - Current activities
  - Code repositories
  - Support
  - Previous WP coordinators

Work package	Coordinator(s)	Mailing list	Mattermost
Overall coordination	Nicole Skidmore Chris Burr (deputy)		
WP1 - Sprucing	Alessandro Bertolin	lhcb-dpa-wp1	link
WP2 - Analysis Productions	Chris Burr	lhcb-dpa-wp2	link
WP3 - Offline Analysis Tools	Mark Smith	lhcb-dpa-wp3	link
WP4 - Innovative Analysis Techniques	Jonas Eschle	lhcb-dpa-wp4	link
WP5 - Legacy Software & Data	Nathan Allen Grieser	lhcb-dpa-wp5	Stripping, DaVinci
WP6 - Analysis Preservation & Open Data	Dillon Fitzgerald	lhcb-data-preservation	link
WP7 - Training and Documentation	Andy Morris	lhcb-dpa-wp7	link

LHCb Data & Analysis Preservation

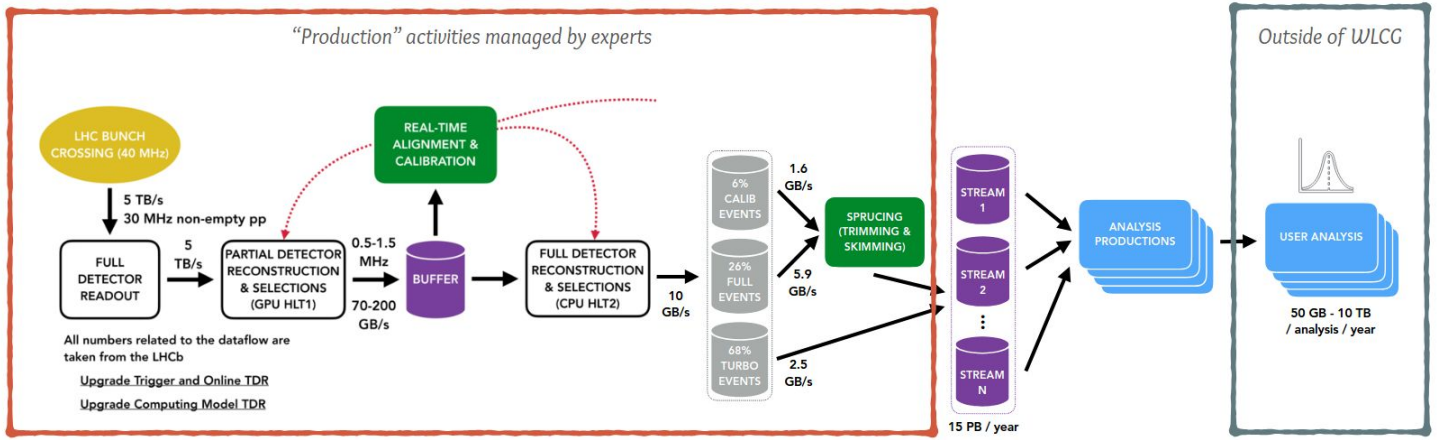
Analysis Productions

Analysis productions provides a central way of producing Ntuples from DSTs

- Preserves a provenance trace from centrally produced DSTs → user Ntuples for analysis
- Includes tool **apd** (**a**nalysis **p**roductions **d**ata) for tagging datasets and facilitating data queries based on meaningful tags

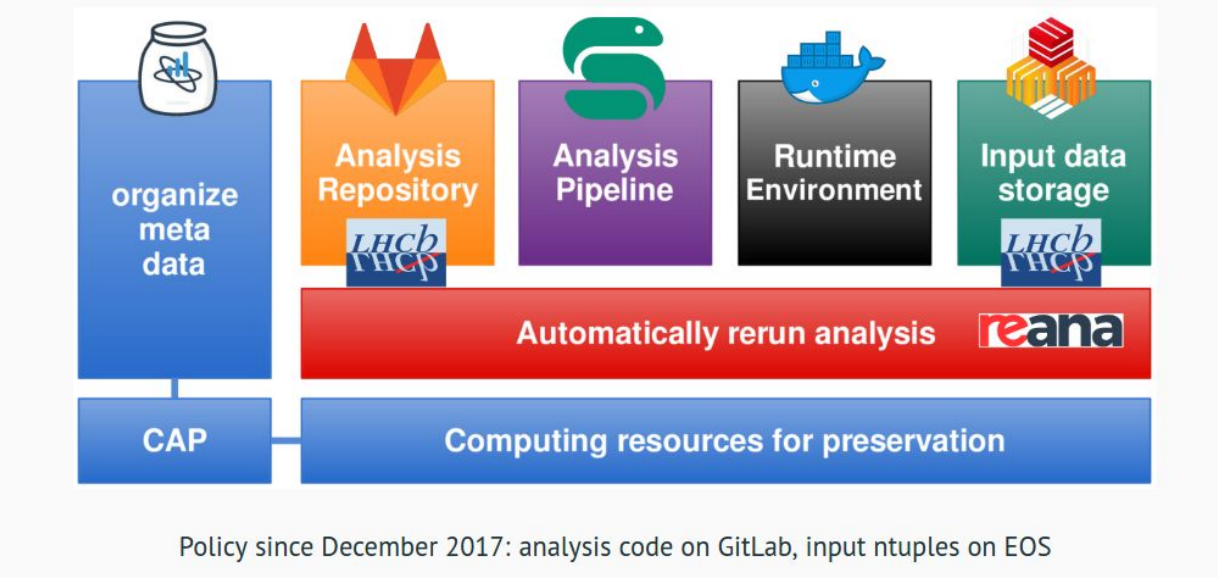
More details

- Analysis Productions (CHEP 2023 presentation)
- apd (CHEP 2023 presentation, CHEP 2023 proceedings)



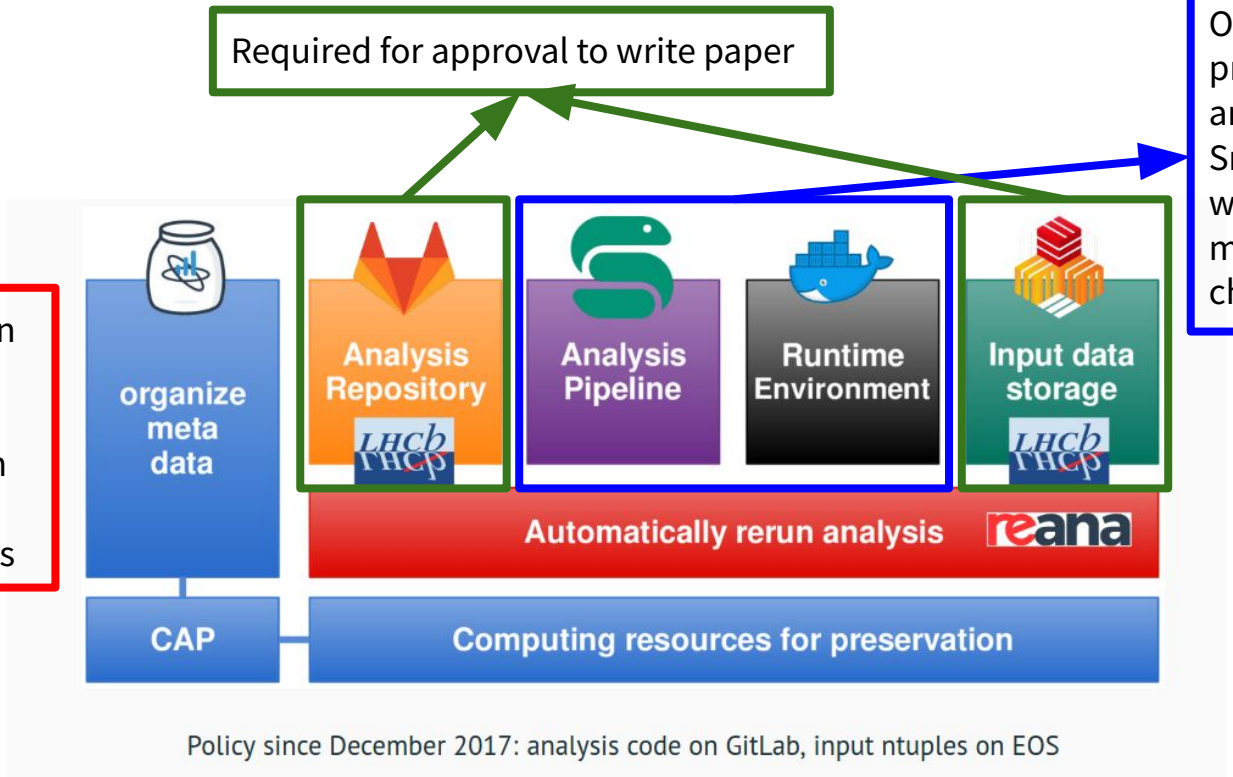
LHCb Analysis Preservation Schematic

From the LHCb Analysis Roadmap (2017)



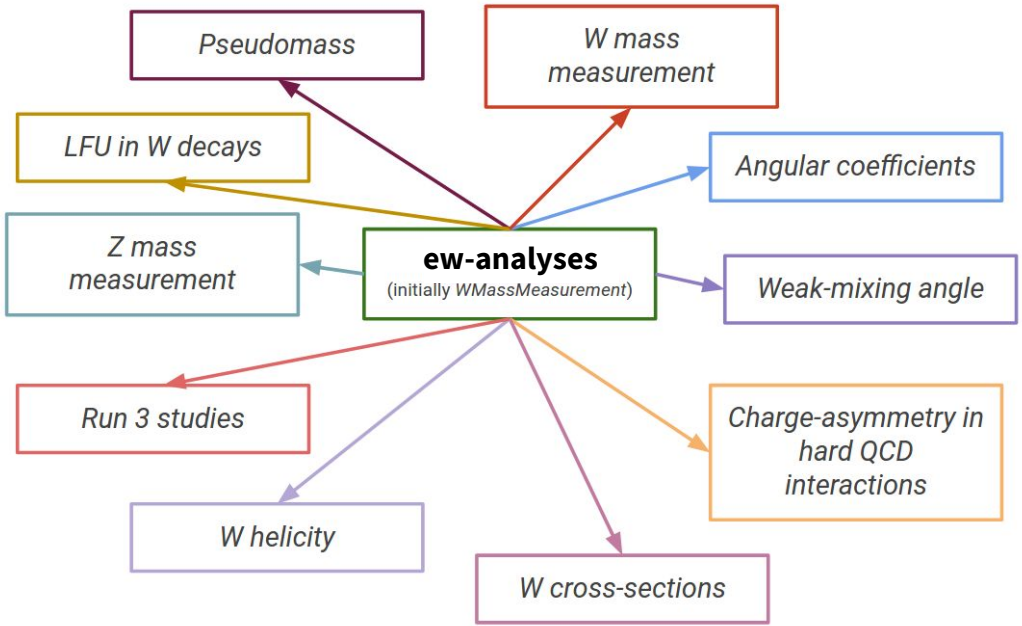
Policy since December 2017: analysis code on GitLab, input ntuples on EOS

LHCb Analysis Preservation Schematic



LHCb Analysis Preservation in Action!

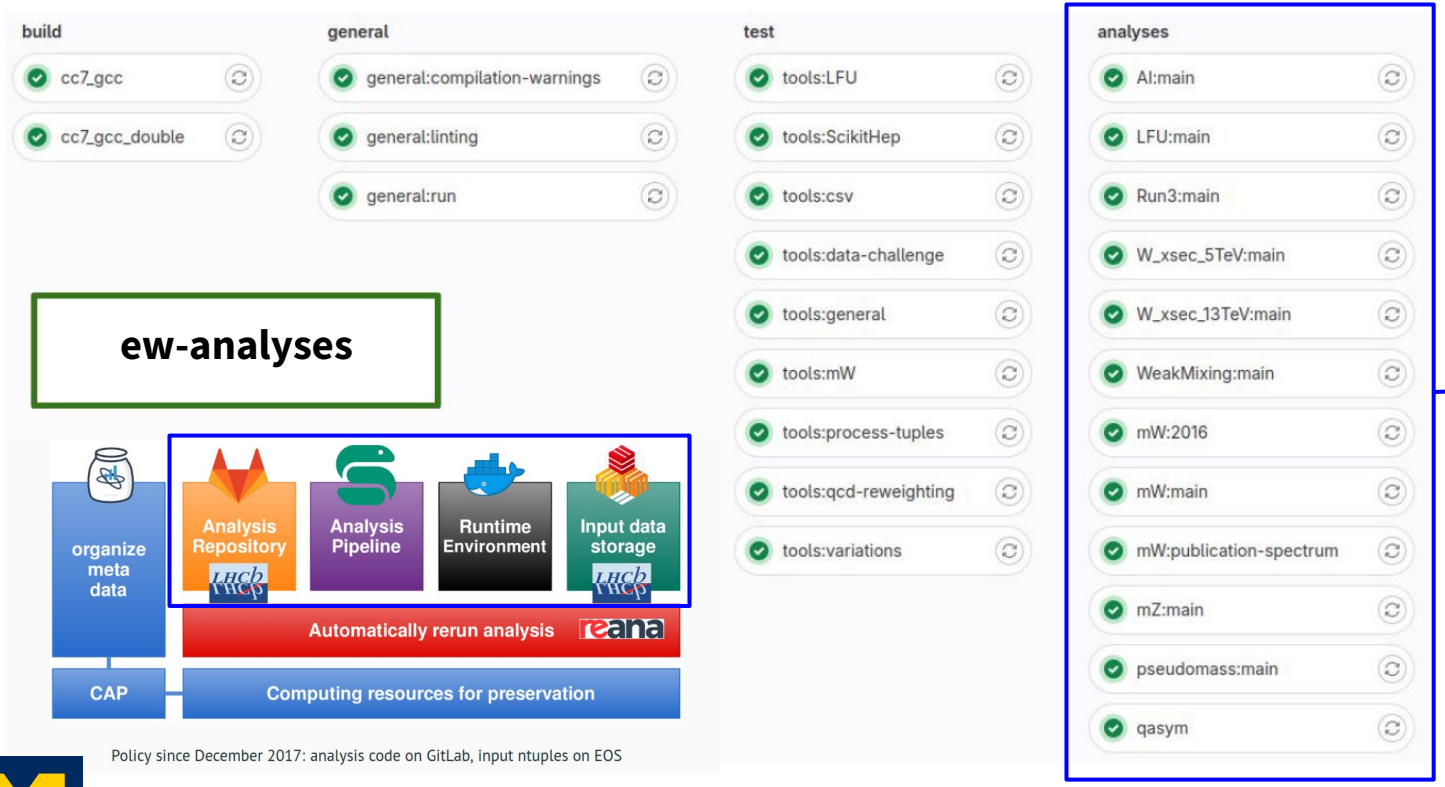
Shared repository for electroweak analyses -- excellent example of LHCb analysis preservation best practices



Born from the W mass measurement

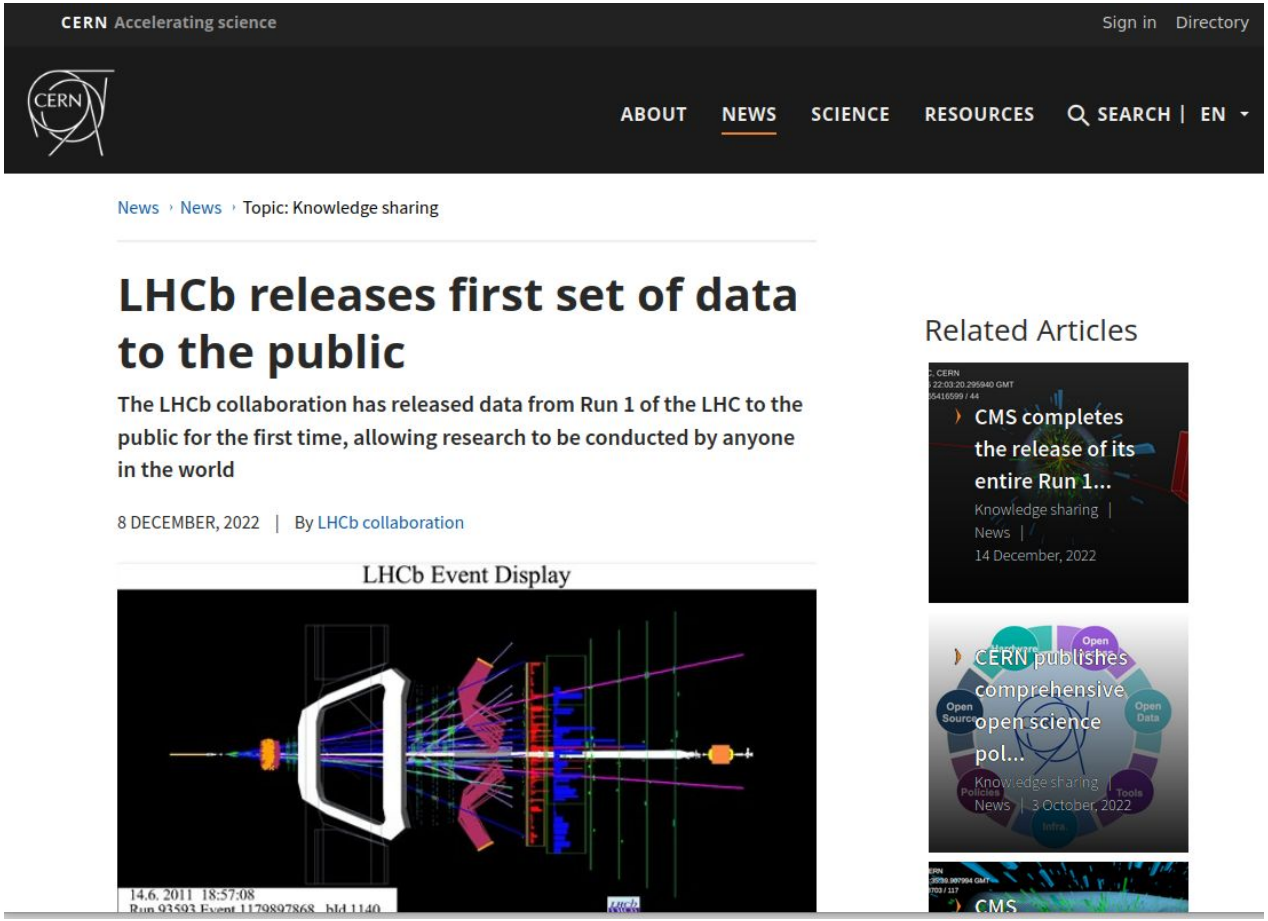
- Naturally evolved into an ecosystem to host related measurements and ensure reproducibility
- Naturally converged on recommended practices for LHCb analysis preservation!

Tracking Changes with Gitlab Continuous Integration



Tracking changes to analysis results when new commits are added! Changes to the final results cause the pipeline to fail -- analyzer can decide to update the reference if changes are intended, or implement a fix

LHCb Open Data



https://home.cern/news/news/knowledge-sharing/lhcb-releases-first-set-data-public

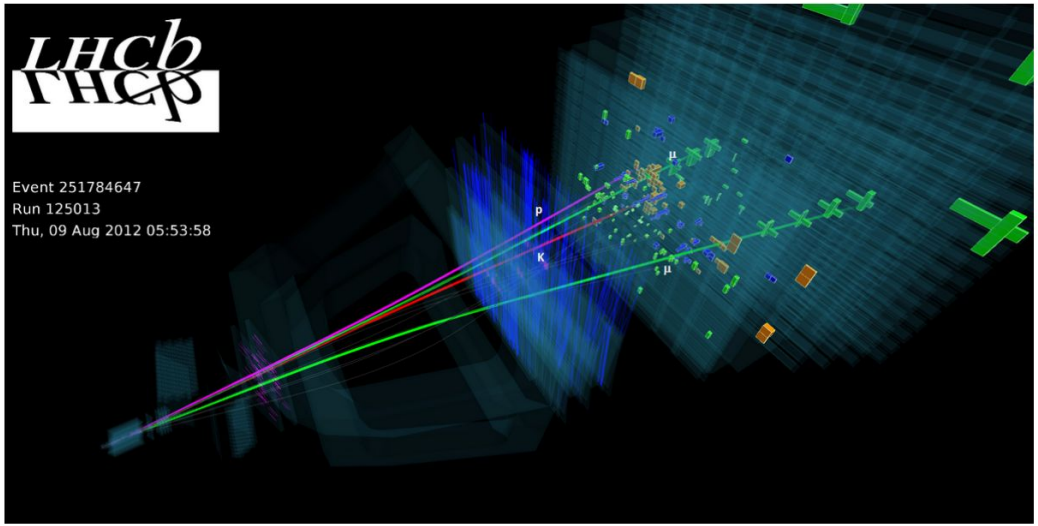
LHCb releases the entire Run I dataset

LHCb Collaboration

News

Date of publication: 2023-12-20

Today the LHCb collaboration completes the release of the data collected throughout the Run I of the Large Hadron Collider at CERN. The sample made available amounts to approximately 800 terabytes (TB) of data. These data, collected by the LHCb experiment in 2011 and 2012, contains information obtained from proton-proton collisions. The format made available provides pre-filtered data, suitable for a wide range of physics studies. The image below displays an event recorded during 2012.



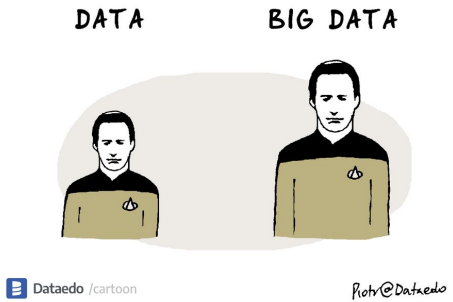
https://opendata.cern.ch/docs/lhcb-releases-entire-run1-dataset

LHCb Open Data

LHCb recently released all of the Run 1 data (~1 PB) on the CERN Open Data Portal:
https://opendata.cern.ch/search?page=1&size=20&experiment=LHCb

Releases for Run 2 and beyond will be challenging due to the volume of data…

	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**



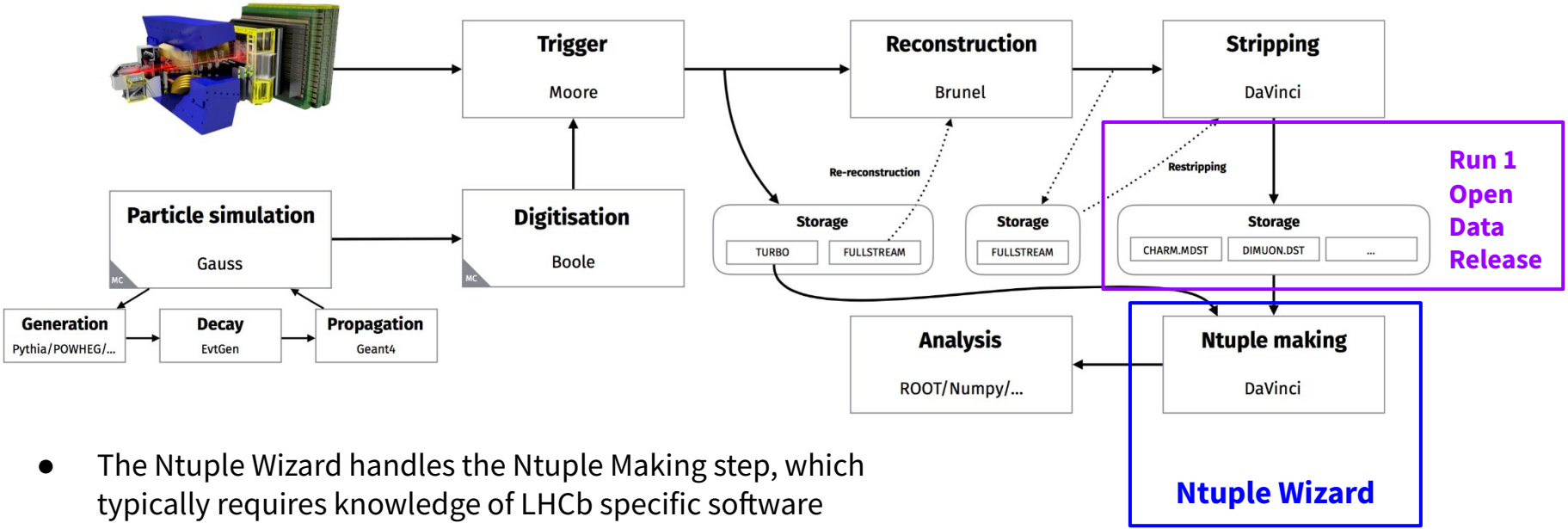
This is not scalable! This prompted the development of a new system…

The LHCb Ntuple Wizard

Previous presentations

- CHEP 2023: Presented idea and mature prototype (screenshots) publicly

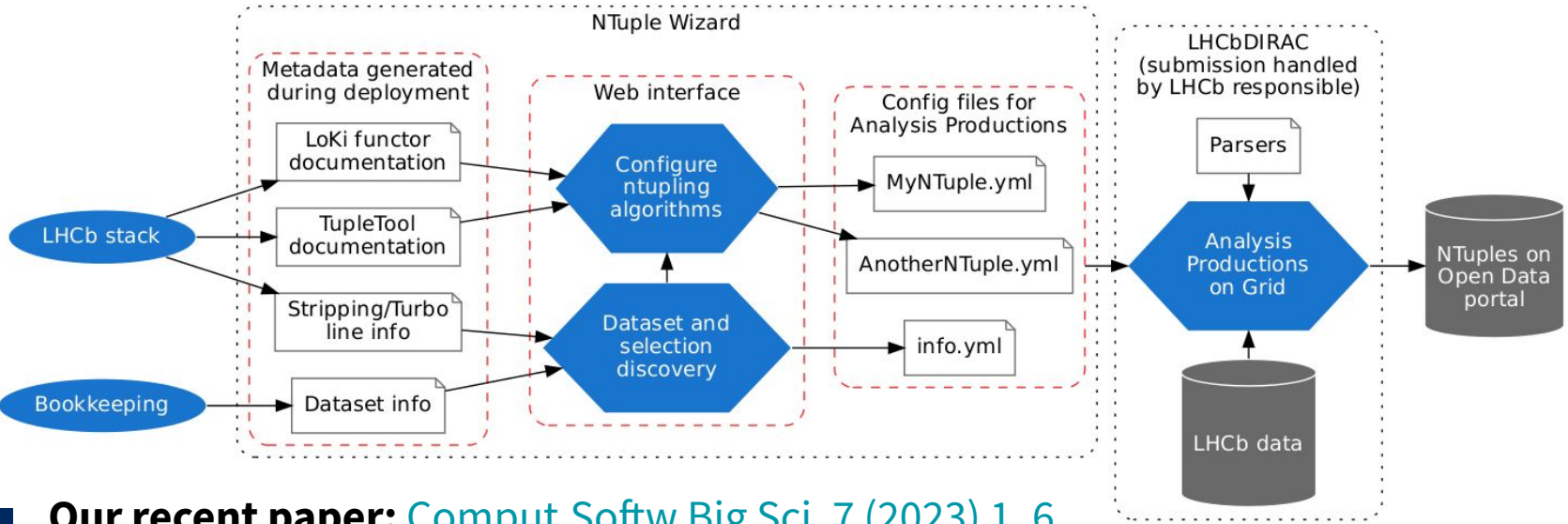
LHCb Run 1 and 2 Data Flow



- The Ntuple Wizard handles the Ntuple Making step, which typically requires knowledge of LHCb specific software
  - **Lower barrier of entry for external analysts!**
  - **Convenient means of navigating available datasets and documentation for internal analysts!**

The Ntuple Wizard

- Intuitive web interface (wizard) guides the user through formulating a query, key features include:
  - Dataset discovery/selection
  - Ntuple configuration
- Input (metadata/documentation) and output (configuration files from user) have secure design features

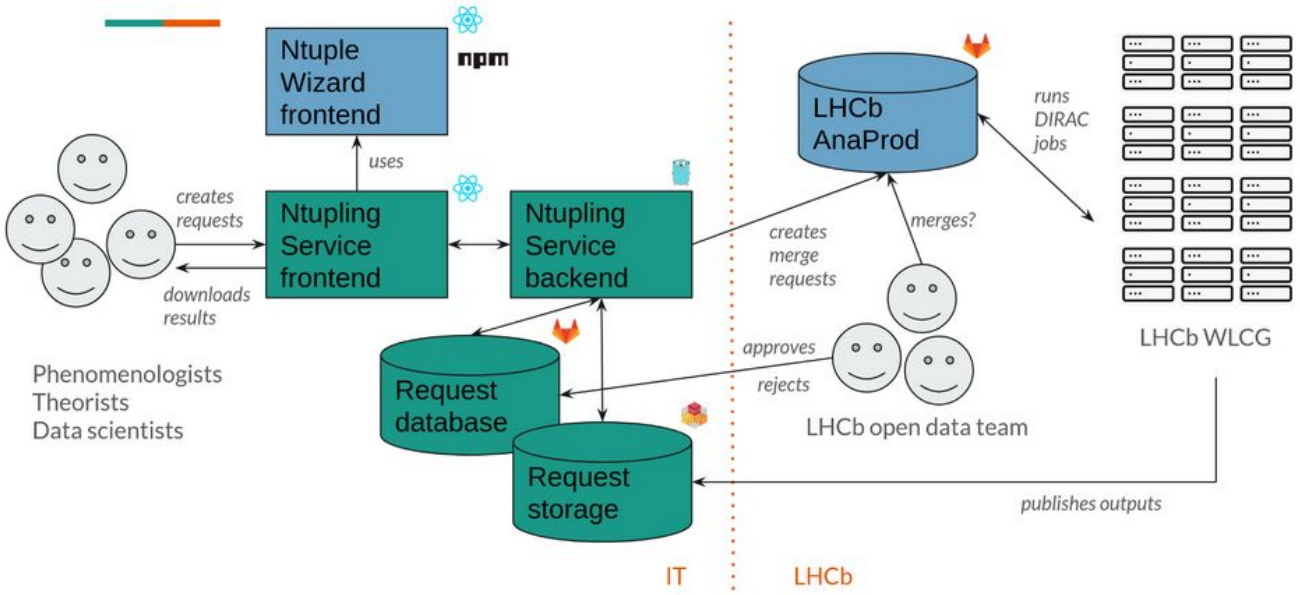


Our recent paper: [Comput.Softw.Big Sci. 7 (2023) 1, 6]

The LHCb Ntupling Service

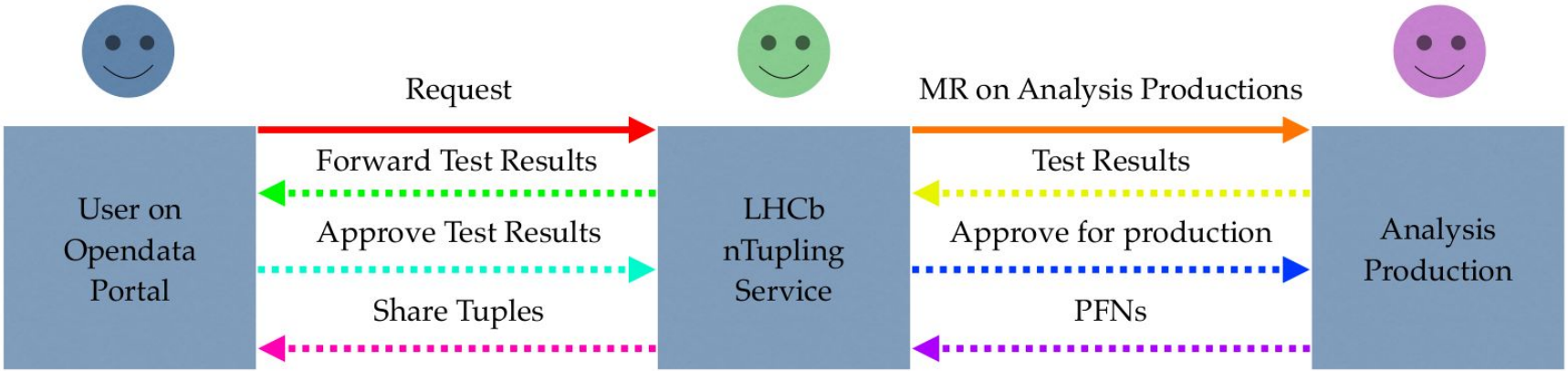
The LHCb Ntuple Wizard has been integrated with the CERN Open Data Portal and the LHCb Analysis Productions batch processing system thanks to the help of the Open Data team from CERN IT. The combined application is called the LHCb Ntupling Service.

- See upcoming talk at CHEP 2024 for more details!



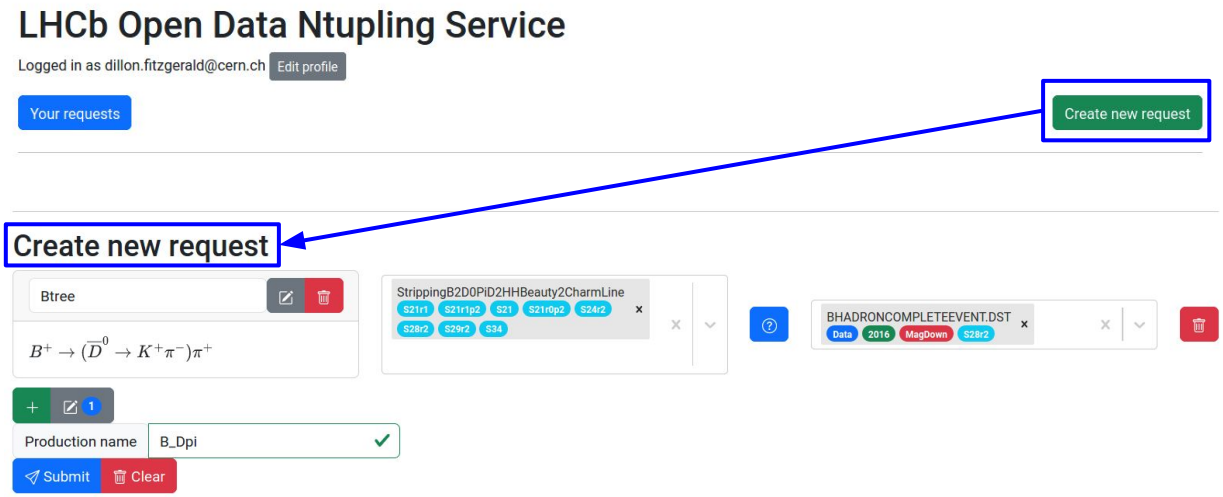
Request Workflow

The LHCb Ntupling Service



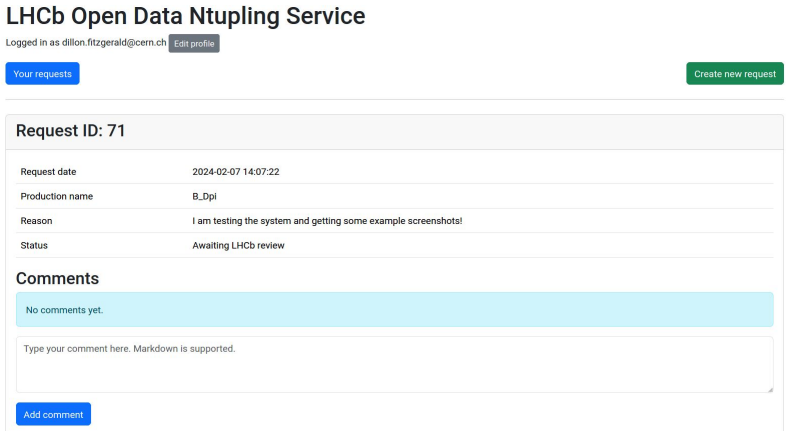
Request

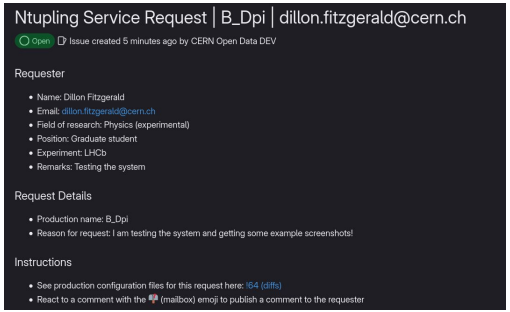
- A request can be made on the LHCb Ntupling Service, to be accessible from the CERN Open Data Portal for the public release
- Create custom Ntuples!
  - Choose from available decays, and data streams
  - Configure your own Ntuple -- Add and customize tools specifying quantities written to output
- Submitting a request will open an issue on GitLab
  - Users will be regularly informed when stages of the production process change

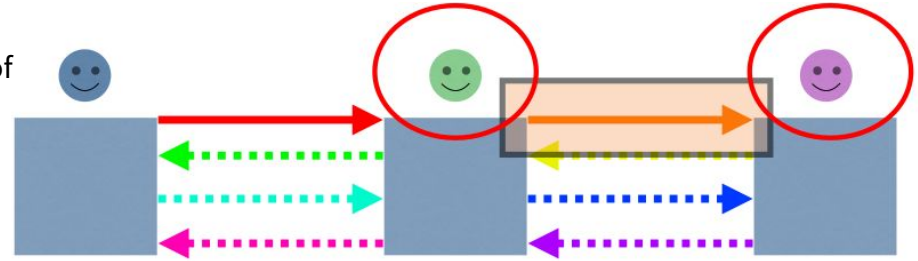


Approval for Test Production by LHCb Open Data Team

- LHCb reviews and approves requests manually
- Approval triggers a draft merge request (MR) in Analysis Productions
    - This triggers a test production via the Analysis Productions pipeline -- metrics from the output are routed to the user
- Merge request left on draft until user confirms results of test production

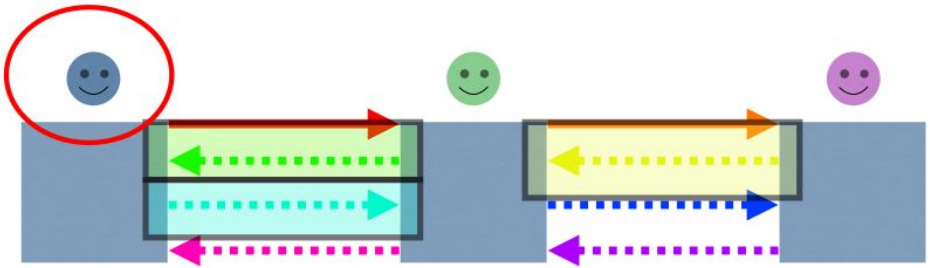






Return Test Results to User for Verification

- From the test production output, a markdown file is provided to the user with information about the production
- The user can verify the test results at this stage
  - Are all expected branches present in the Ntuple?
  - Is the estimated output size within a feasible range?
- After user approval, the merge request on Analysis Productions is marked as ready



test-production.md

Welcome to the **LHCb** Open Data service!

Congratulations! Your request has been approved by the *LHCb Open Data Team* and submitted to the *LHCb Analysis Productions service*!

The internal testing system has been successfully completed, creating a small test sample.

Please take a moment to assess the test statistic below and consider the following:

1. Is the estimated output size within a feasible range for you?
2. Do all expected variables exist in the ROOT file?

If the test statistic fulfills the above criteria and you are satisfied with the results, please confirm your request here.

Next Steps:

- Once you confirm you are satisfied with the test result, the **LHCb** Open Data Team will send the request to full production
- You will be notified once the production is finished. The ROOT files are then transfered to the Open Data Portal.
- You can then access the data freely. Have fun on your Analysis!

Thank you for your patience. If you have further questions or concerns, please feel free to reach out by leaving a comment on your request, by sending an email to opendata-support@cern.ch or by opening a ticket on the Opendata Data Forum.

`B_JPsiK` will process 135.6TB of data and create about 17.8GB across 6 samples. Please note that the actual size of the output files may vary from the estimate.

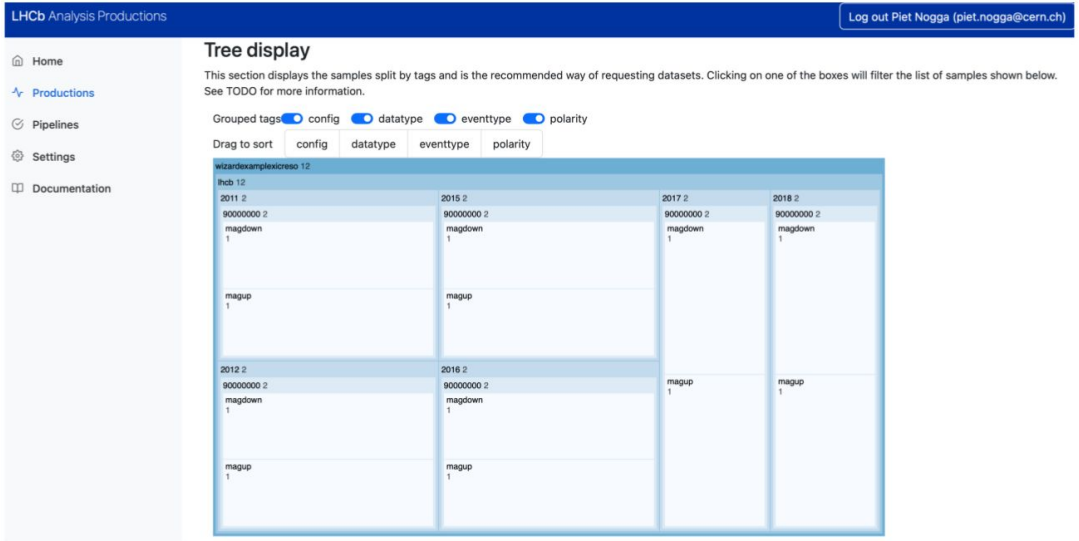
▼ **See Branches in Production Tuples**
▼ job5
▼ Btree/DecayTree;1 (394 Branches)
Bplus_ENDVERTEX_X
Bplus_ENDVERTEX_Y
Bplus_ENDVERTEX_Z
Bplus_ENDVERTEX_XERR
Bplus_ENDVERTEX_YERR
Bplus_ENDVERTEX_ZERR
Bplus_ENDVERTEX_CHI2
Bplus_ENDVERTEX_NDOF
Bplus_ENDVERTEX_COV_
Bplus_OWNPV_X
Bplus_OWNPV_Y
Bplus_OWNPV_Z
Bplus_OWNPV_XERR
Bplus_OWNPV_YERR
Bplus_OWNPV_ZERR
Bplus_OWNPV_CHI2
Bplus_OWNPV_NDOF
Bplus_OWNPV_COV_
Bplus_IP_OWNPV
Bplus_IPCHI2_OWNPV
Bplus_FD_OWNPV

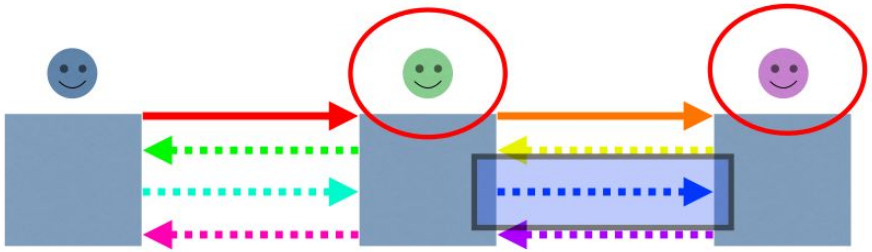
Download Close

Approval for Full Production

LHCb approves the final production and merges to Analysis Productions

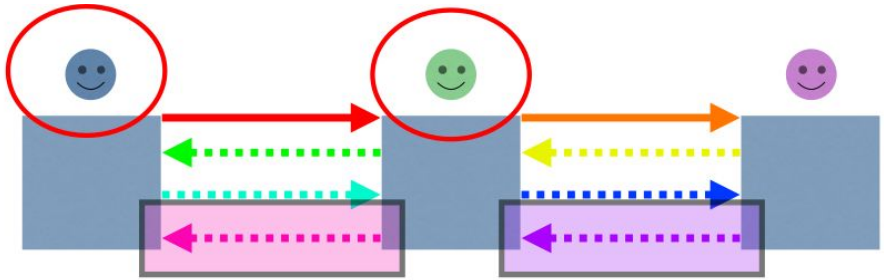
- This triggers the Analysis Productions system to process the request over all selected datasets





Deliver Ntuples to CERN Open Data Portal

- Upon completion, the output Ntuples are copied to *eospublic* and hosted on the CERN Open Data Portal
- The produced Ntuples can now be downloaded by the user!



Production Output (10)

PFNs

```
root://eoslhcb.cern.ch//eos/lhcb/grid/prod/lhcb/LHCb/Collision15/DVNTUPLE.ROOT/00210598/0000/00210598_00000004_1.dvntuple.root
root://eoslhcb.cern.ch//eos/lhcb/grid/prod/lhcb/LHCb/Collision15/DVNTUPLE.ROOT/00210598/0000/00210598_00000002_1.dvntuple.root
root://eoslhcb.cern.ch//eos/lhcb/grid/prod/lhcb/LHCb/Collision15/DVNTUPLE.ROOT/00210598/0000/00210598_00000003_1.dvntuple.root
root://eoslhcb.cern.ch//eos/lhcb/grid/prod/lhcb/LHCb/Collision15/DVNTUPLE.ROOT/00210598/0000/00210598_00000005_1.dvntuple.root
```

The real production for your request has been completed! You can download the results below.

Real production results

Data	Size	Preview
b_jpsik_2016_magdown_0.root	1.409 GiB	
b_jpsik_2016_magdown_1.root	559.563 MiB	
b_jpsik_2016_magup_0.root	1.404 GiB	
b_jpsik_2016_magup_1.root	507.224 MiB	
b_jpsik_2017_magdown_0.root	1.424 GiB	
b_jpsik_2017_magdown_1.root	568.706 MiB	
b_jpsik_2017_magup_0.root	1.428 GiB	
b_jpsik_2017_magup_1.root	482.238 MiB	
b_jpsik_2018_magdown_0.root	1.349 GiB	
b_jpsik_2018_magdown_1.root	940.678 MiB	
b_jpsik_2018_magup_0.root	1.447 GiB	
b_jpsik_2018_magup_1.root	1.015 GiB	

Alpha Release → Beta Release

The LHCb Ntuple Wizard & Ntupling Service were officially released to the LHCb collaboration in February (alpha release)

- The goal was to get useful feedback from collaborators before opening more broadly to some affiliated theorists for the beta release.

The beta release will proceed at the [First LHCb Open Data and Ntuple Wizard Workshop] on 22 October, 2024

Summary

- Many ongoing activities related to analysis preservation and open data at LHCb

- Analysis Preservation
  - Analysis preservation best practices being adopted by many analysts

- Open Data
  - Full Run 1 dataset released to the CERN Open Data Portal in December 2023
  - Novel application for Ntuple creation from LHCb data using a web interface is in the mature stages of development -- the LHCb Ntuple Wizard & Ntupling Service
    - Access to run 2 data!
    - Beta release planned for 22 October, 2024, at the First LHCb Open Data and Ntuple Wizard Workshop

BACKUP

Analysis Life Cycle Management Tool (ALCM)

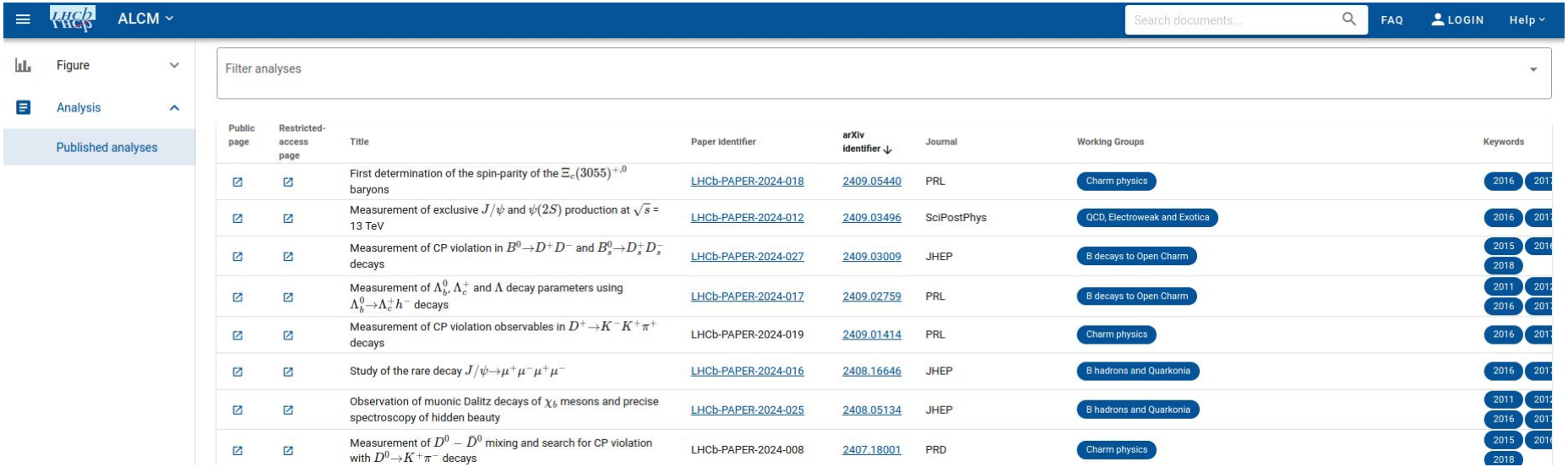
ALCM is a database tool

- Organization of the review workflow
  - Workflow tracker connected to membership database
- Store relevant information for analyses
  - Enhance interpretability of results
- Unification and replacement of old working group databases, editorial board database, and public pages for results
  - Centralize important information related to physics results

Previous slides covered preservation of analysis code and workflows, while ALCM is for preservation of the review process towards publication

Analysis Life Cycle Management Tool (ALCM)

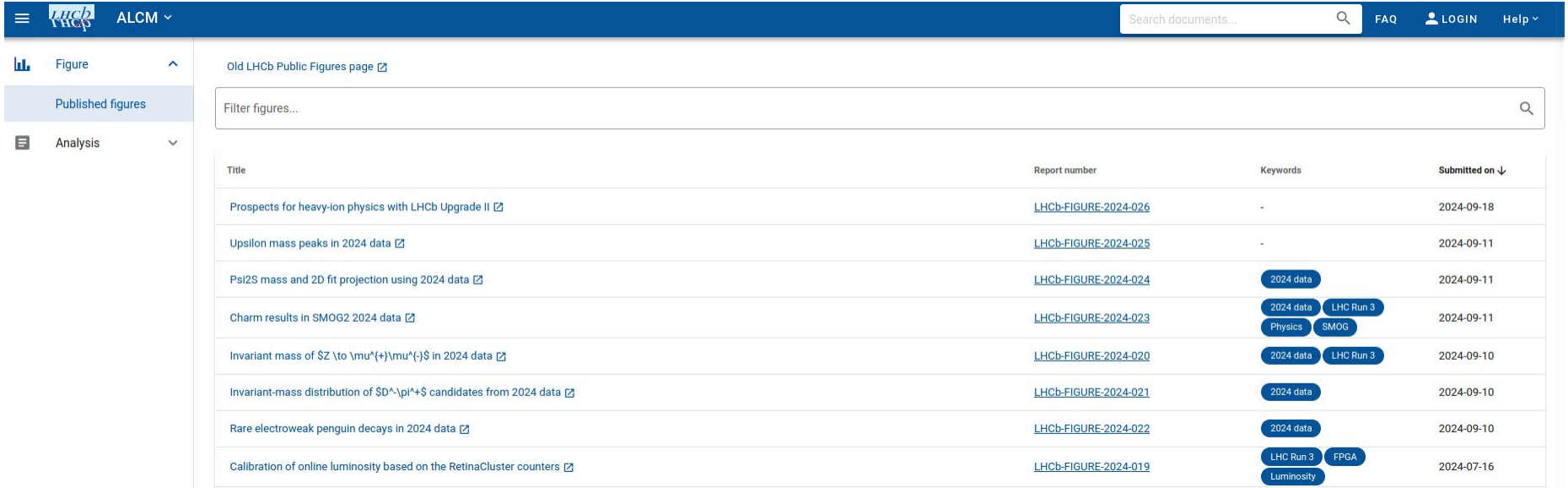
Published analyses



Public page	Restricted-access page	Title	Paper identifier	arXiv identifier	Journal	Working Groups	Keywords
		First determination of the spin-parity of the $\Xi_c(3055)^{+,0}$ baryons	LHCb-PAPER-2024-018	2409.05440	PRL	Charm physics	2016 2017
		Measurement of exclusive J/ψ and $\psi(2S)$ production at $\sqrt{s}$ = 13 TeV	LHCb-PAPER-2024-012	2409.03496	SciPostPhys	QCD, Electroweak and Exotica	2016 2017
		Measurement of CP violation in $B^0 \to D^+ D^-$ and $B_s^0 \to D_s^+ D_s^-$ decays	LHCb-PAPER-2024-027	2409.03009	JHEP	B decays to Open Charm	2015 2016 2018
		Measurement of Λ_b^0, Λ_c^+ and Λ decay parameters using $\Lambda_b^0 \to \Lambda_c^+ h^-$ decays	LHCb-PAPER-2024-017	2409.02759	PRL	B decays to Open Charm	2011 2012 2016 2017
		Measurement of CP violation observables in $D^+ \to K^- K^+ \pi^+$ decays	LHCb-PAPER-2024-019	2409.01414	PRL	Charm physics	2016 2017
		Study of the rare decay $J/\psi \to \mu^+ \mu^- \mu^+ \mu^-$	LHCb-PAPER-2024-016	2408.16646	JHEP	B hadrons and Quarkonia	2016 2017
		Observation of muonic Dalitz decays of χ_b mesons and precise spectroscopy of hidden beauty	LHCb-PAPER-2024-025	2408.05134	JHEP	B hadrons and Quarkonia	2011 2012 2016 2017
		Measurement of $D^0 - \bar{D}^0$ mixing and search for CP violation with $D^0 \to K^+ \pi^-$ decays	LHCb-PAPER-2024-008	2407.18001	PRD	Charm physics	2015 2016 2018

Analysis Life Cycle Management Tool (ALCM)

Published figures



Analysis Life Cycle Management Tool (ALCM)

Analysis entry

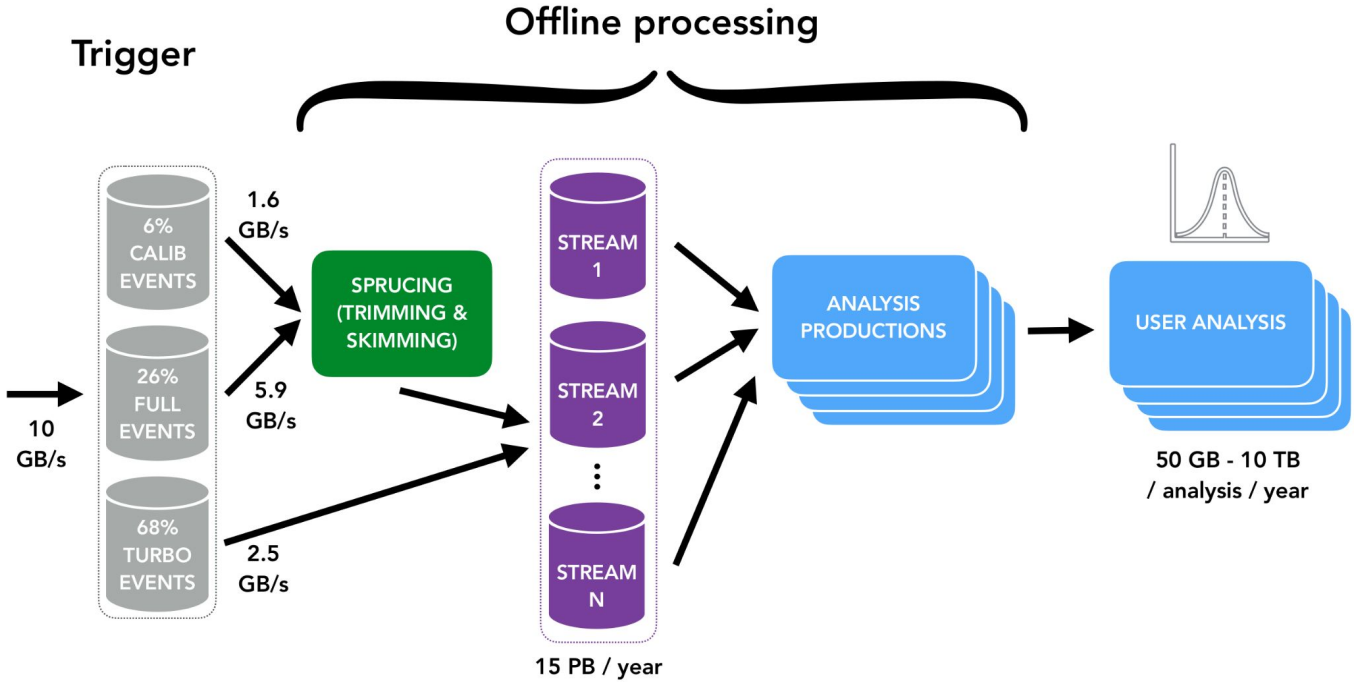
Analysis Summary for "Measurement of CP violation in $B^0 \to D^+ D^-$ and $B_s^0 \to D_s^+ D_s^-$ decays"

Property	Value
Paper	LHCb-PAPER-2024-027
Title	Measurement of CP violation in $B^0 \to D^+ D^-$ and $B_s^0 \to D_s^+ D_s^-$ decays
Hashtags	2015 2016 2017 2018
Physics working groups	B decays to Open Charm
EP Number	CERN-EP-2024-217
arXiv	2409.03009
Journal	JHEP
Additional information (only available for LHCb members)	Go to restricted access page
Abstract	A time-dependent, flavour-tagged measurement of \CP violation is performed with \decay{\Bd}{\Dp\Dm} and \decay{\Bs}{\Dsp\Dsm} decays, using data collected by the \lhcb detector in proton-proton collisions at a centre-of-mass energy of 13\tev corresponding to an integrated luminosity of 6\invfb. In \decay{\Bd}{\Dp\Dm} decays the \CP-violation parameters are measured to be $S_{\mathrm{D^+D^-}} = \SResultRunTwo$, $C_{\mathrm{D^+D^-}} = \CResultRunTwo$. In \decay{\Bs}{\Dsp\Dsm} decays the \CP-violating parameter formulation in terms of \phis and $\lvert\lambda\rvert$ results in $\phis = \phisResultRunTwo$, $\lvert\lambda_{\mathrm{D_s^+D_s^-}}\rvert = \lambdaResultRunTwo$. These results represent the most precise single measurement of the \CP-violation parameters in their respective channels. For the first time in a single measurement, \CP symmetry is observed to be violated in \decay{\Bd}{\Dp\Dm} decays with a significance exceeding six standard deviations.

Figures and captions

DOWNLOAD PLOTS

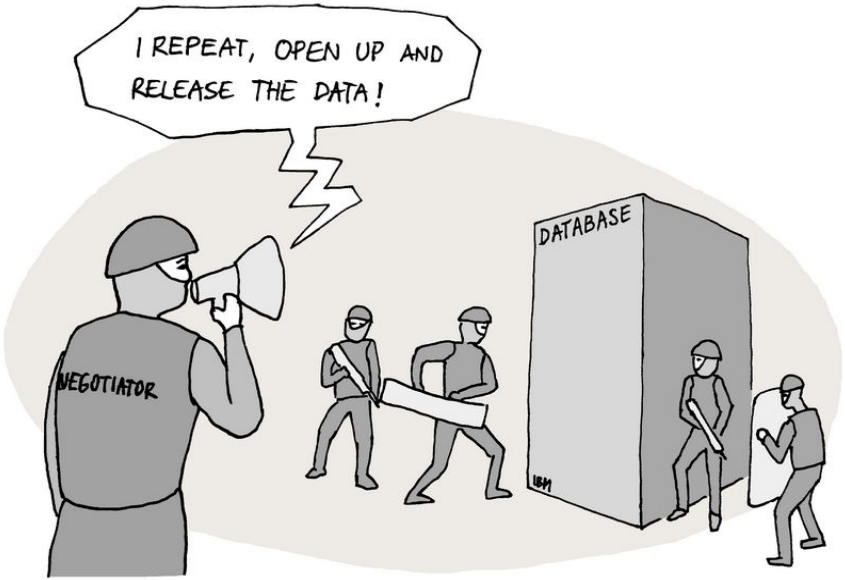
Run 3 Data Flow



LHCb-FIGURE-2020-016

Open Data

CMS Open Data User Story: The Future of Particle Physics is Open [2017-12-01 by Jesse Thaler (MIT)]
(includes links to 2 published papers with open data!)



Open Data

The data collected at the LHC is very valuable! It should be made available to the public in accordance with the CERN Open Data Policy and CERN Open Science Policy

- This takes a considerable amount of work. Today I will talk about some of LHCb's efforts to do so

The CERN Open Data Portal (https://opendata.cern.ch/) provides a location for LHC experiments to host open data

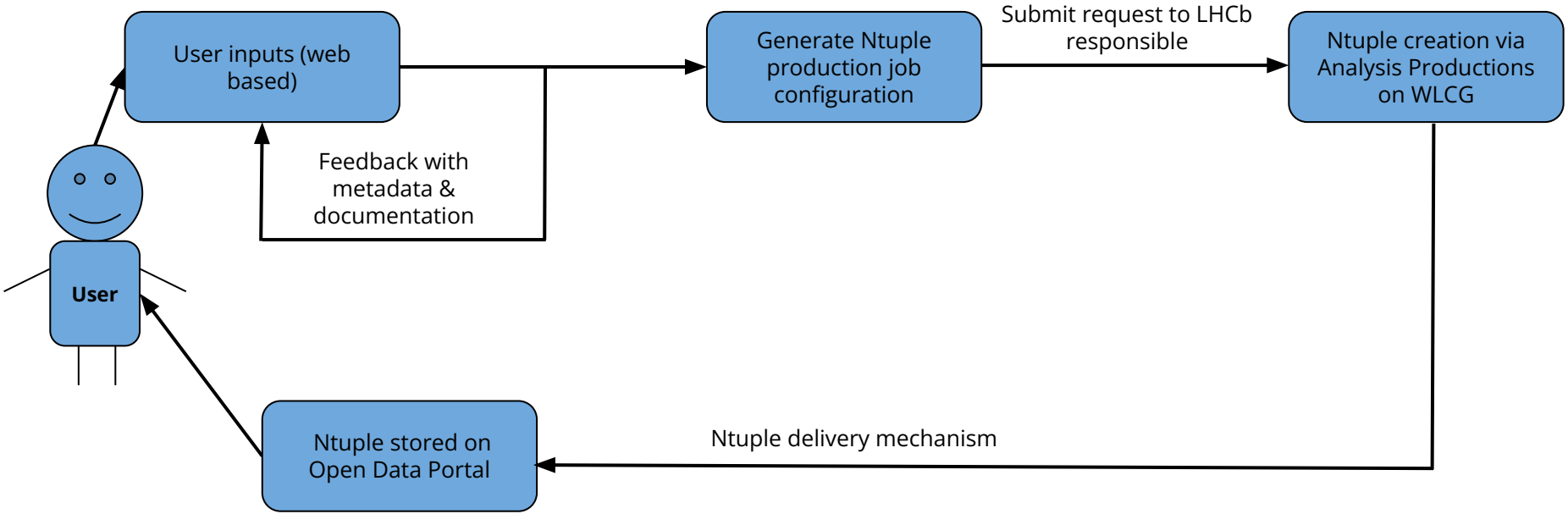
CERN Open Data Policy

The CERN Open Data Policy outlines the commitment to make the data collected at the LHC publicly available at several levels of complexity, as established by the Data Preservation in High Energy Physics Collaboration (DPHEP-2012-001)

- Level 1: Published results
  - This can include tables and figures but also preprocessed Ntuples or binned and unbinned fit likelihood functions.
- Level 2: Outreach and education
  - Usually in the form of highly preprocessed Ntuples.
- Level 3: Reconstructed data
  - These data have been preprocessed to derive physics objects, such as charged particle candidates, photons, or particle jets. Reconstructed data may or may not be corrected for detector effects, such as efficiency and resolution.

  Target: Release research quality data mainly for theorists and phenomenologists
- Level 4: Raw data
  - the basic quantities recorded by the experimental instruments.

The Ntuple Wizard



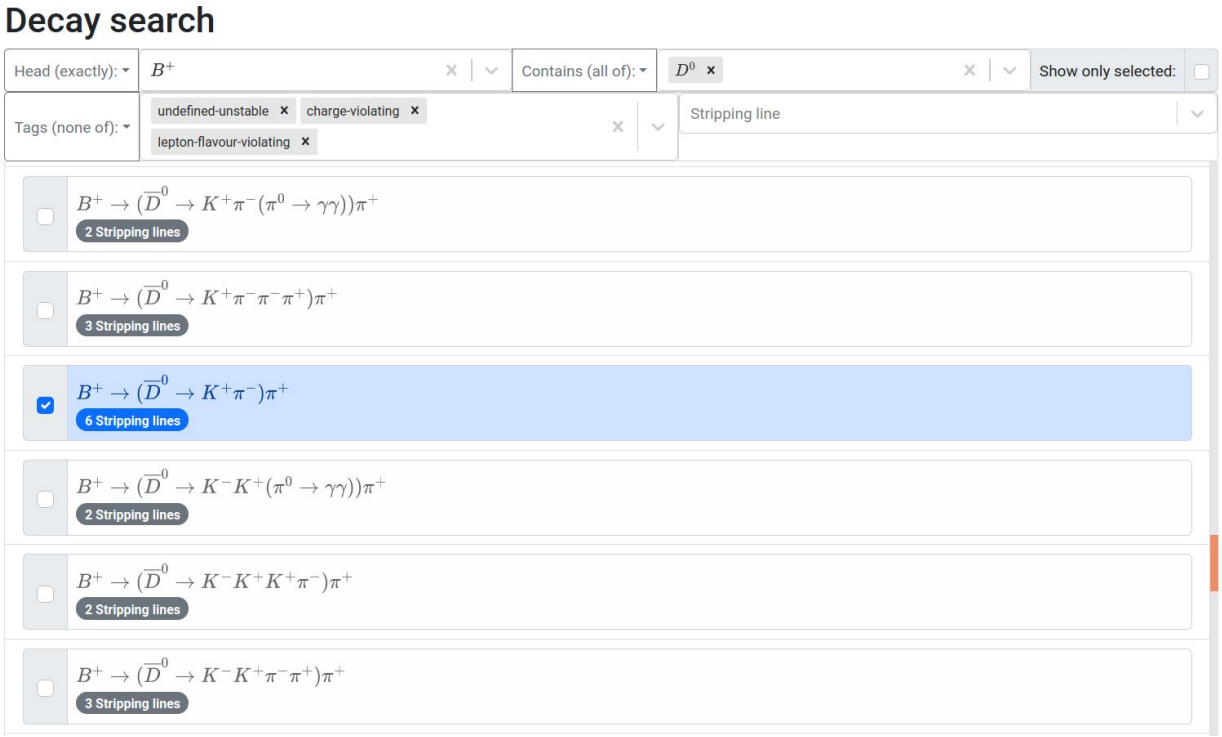
Our recent paper: [Comput.Softw.Big Sci. 7 (2023) 1, 6]

Interlude: Security & Permissions

- Standard LHCb Ntuple making application (DaVinci) configured with python scripts
  - **Running arbitrary code from external users is a security risk!**
  - **Config output saved in yaml data structures, interpreted by internal parsers**
- Dataset discovery and Ntuple configuration require metadata from the LHCb database and software stack
  - Metadata is extracted at “deployment time”
  - **Only static files are read at run time, no interaction with LHCb database system**
- LHCb policy reserves right to withhold part of a dataset (e.g. ongoing analyses)
  - Run 1 open data release initially only contained a subset of the data because of this
  - The Ntuple Wizard can improve this via **fine-grained control** over:
    - **Stripping lines** or equivalent selections

Web Interface: Dataset Discovery

***Key feature:** Find available dataset by first choosing physics object of interest!



Lists physics objects available in the LHCb database (primarily decays)

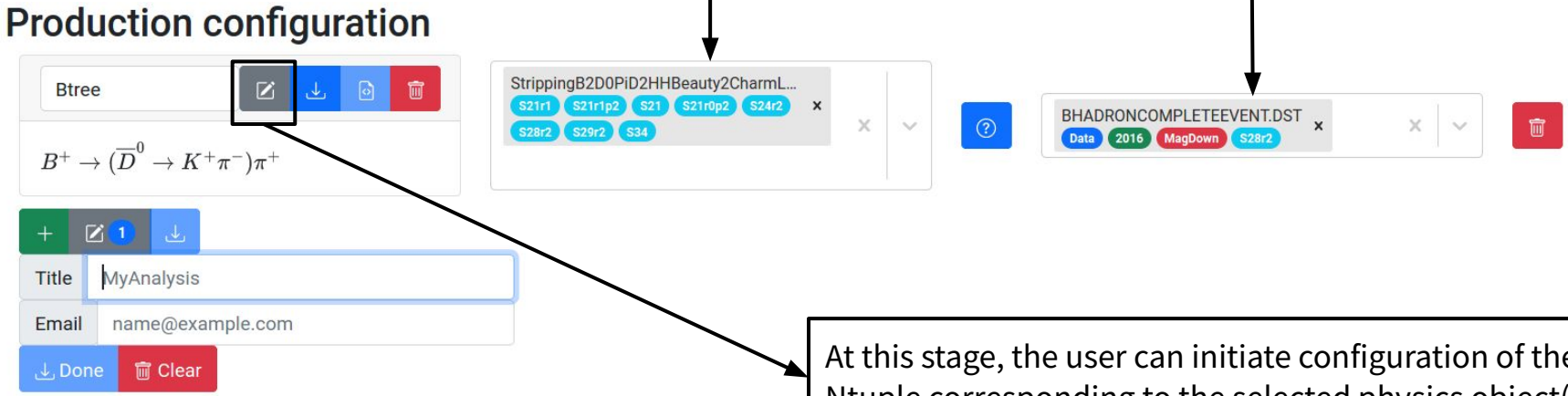
- List filtering options include:
  - Decay head (top level decaying particle)
  - Particles in the decay
  - Tags related to specific physics (include or exclude)
  - “Stripping line” name
    - more useful for LHCb internal users
- Can make multiple selections from the list

Web Interface: Dataset Discovery

Selection of a physics object exposes the corresponding available datasets for the user to choose from

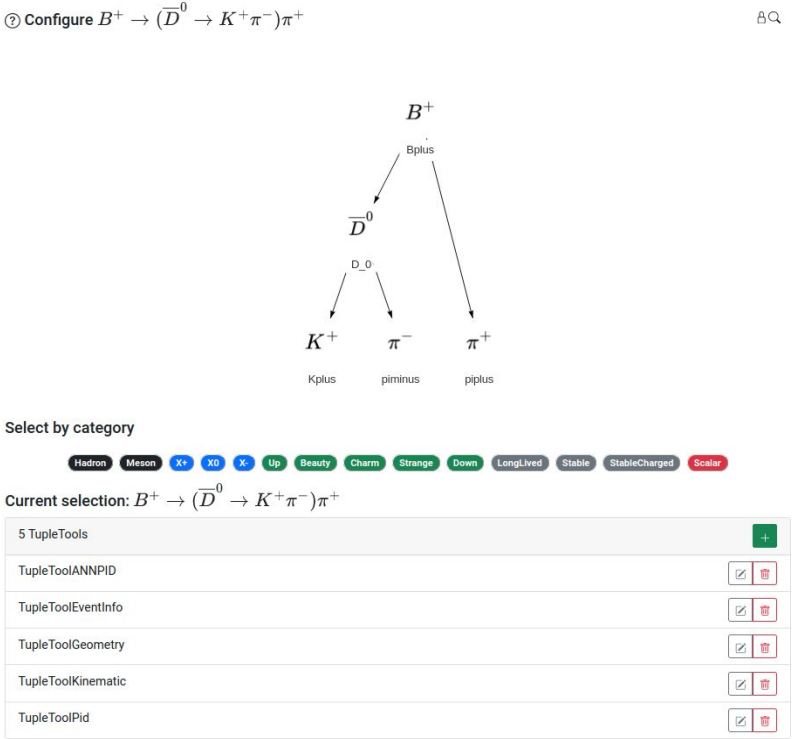
Stripping line selection
- Specifies algorithms applied to identify candidates of the selected physics object

Dataset selection
- Specifies the dataset to run over -- multiple selections can be made



At this stage, the user can initiate configuration of the Ntuple corresponding to the selected physics object(s)

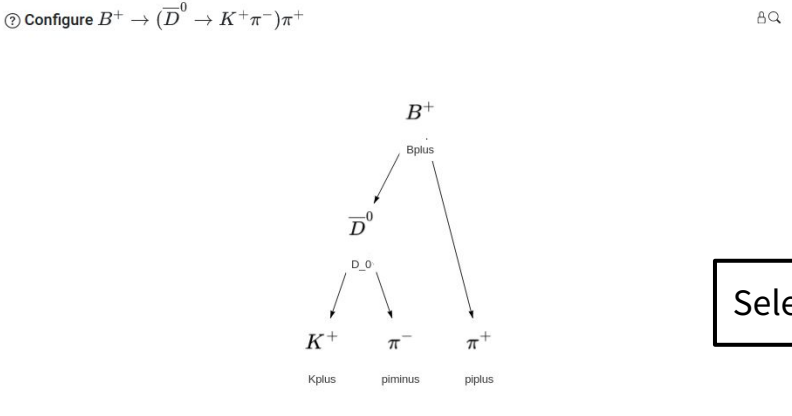
Web Interface: Ntuple Configuration

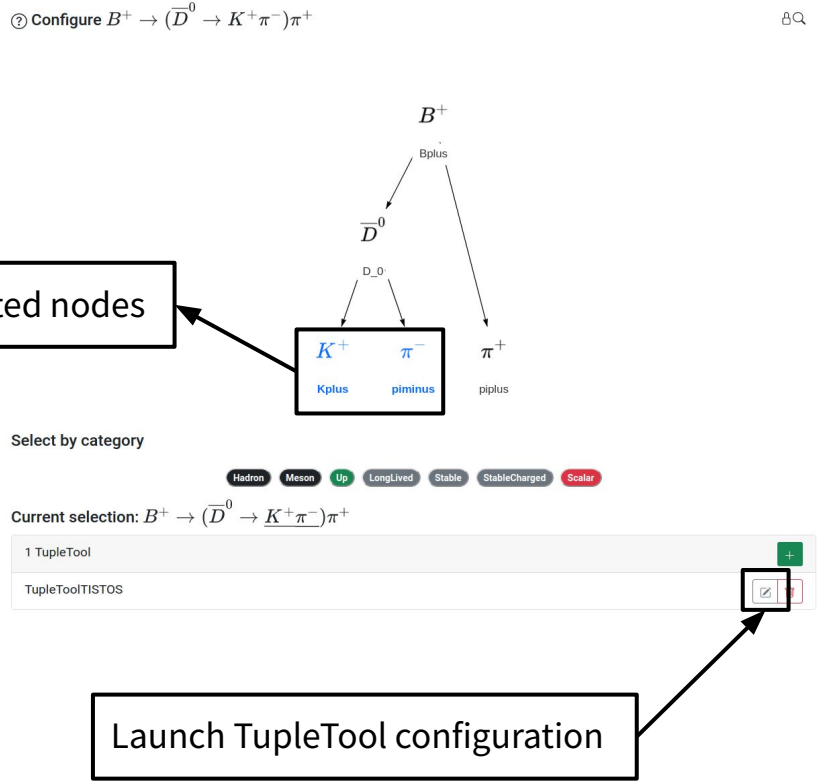


Ntuple configuration via an interactive node tree

- Particles in decay rendered as nodes in tree
- Each node can be configured independently, or in various groupings
  - Labels provided to select nodes by similar categories
- Node configuration proceeds by adding, removing, or configuring **TupleTools**, which save various physics quantities to the Ntuple
  - Can be performed on entire tree, single node, or selection of nodes
- The entire node tree includes 5 standard TupleTools for LHCb analysis by default

Web Interface: Ntuple Configuration

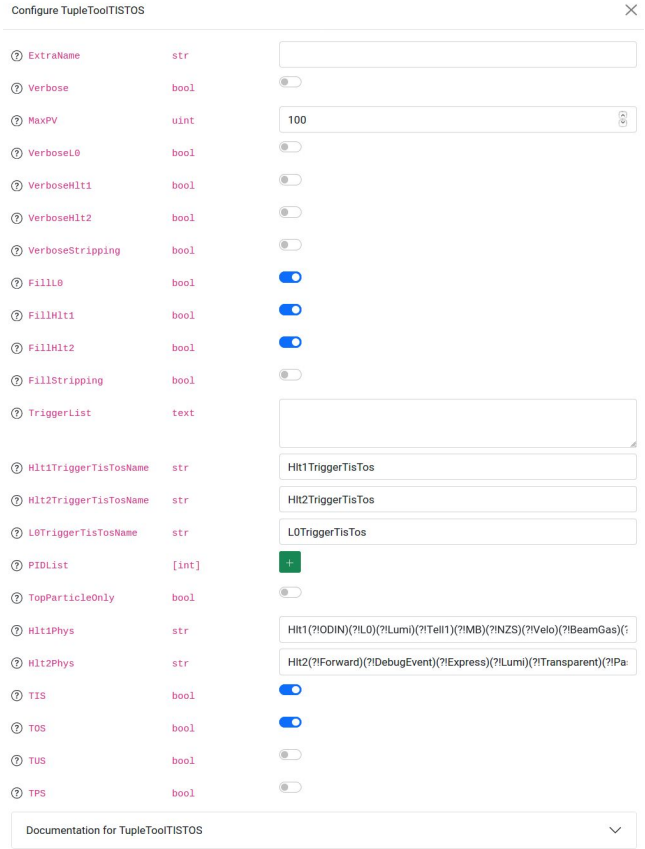




Web Interface: Tuple Tool Configuration Example

Example of TupleTool configuration interface for TupleToolTISTOS (**T**rigger **I**ndependent of **S**ignal/ **T**rigger **o**n **S**ignal)

- Configurable names, data types, and user input fields are included
- Mouseover tooltips and links to documentation are included for guidance
  - This includes LHCb Doxygen documentation
- Each TupleTool has specific configurables
  - For many tools, the standard configuration is perfectly fine
  - Only certain tools (e.g. related to the trigger) need specific configurations, to be specified in the documentation



Happy Analysis!