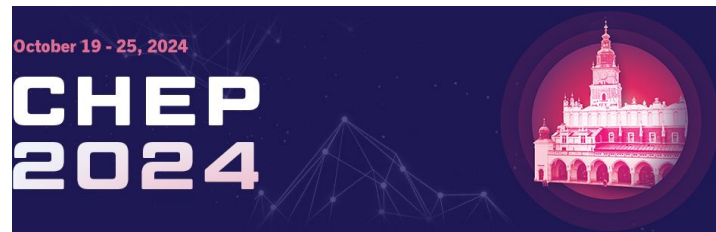


LHCb Open Data Ntupling Service: On-demand production and publishing of custom LHCb Open Data

Christine Aidala¹, Dillon Fitzgerald¹, Kai Habermann², Ludwig Kramer², Adam Morris,³ Sebastian Neubert², **Piet Nogga**², Eduardo Rodrigues⁴, Marco Donadoni³, Daan Rosendal³, Tibor Simko³

¹University of Michigan, ²University of Bonn, ³CERN, ⁴University of Liverpool



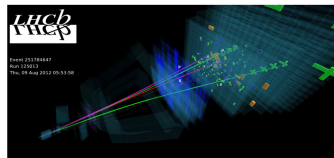
Open Data at LHCb

- ▶ **Cern Open Data** boosts transparency, collaboration and innovation by making experimental data available
- ▶ The **CERN Open Data Policy** encourages the release of reconstructed data <https://cds.cern.ch/record/2745133>
- ▶ The Open Data from all LHC experiments is released through the Open Data Portal (supported and maintained by CERN)
- ▶ LHCb agreed to publish data from a data taking campaign after 5 years (full dataset after 10 years)
- ▶ **Milestone:** Full ~1 Pb Run 1 release in 2023!

LHCb releases the entire Run 1 dataset

2023-10-01 by LHCb Collaboration

Today the LHCb collaboration completes the release of the data collected throughout the Run 1 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>

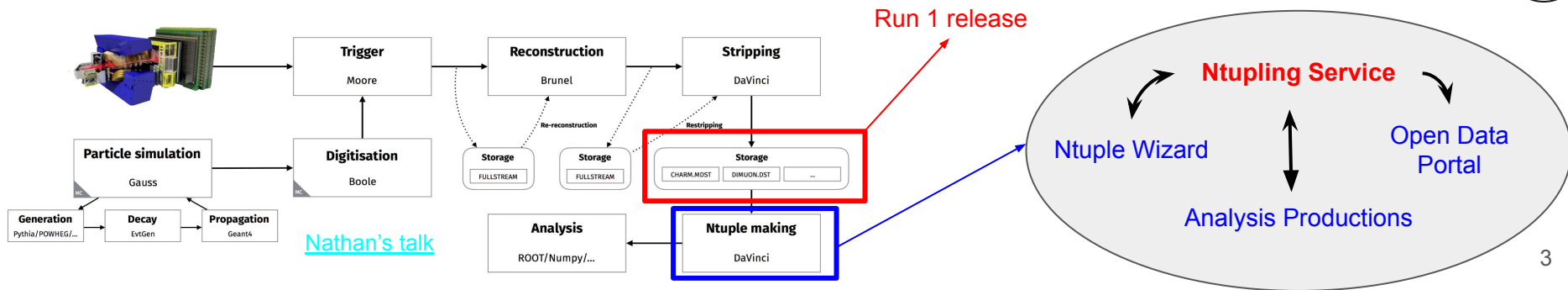
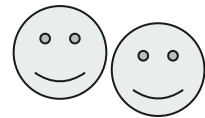
	ALICE	ATLAS	CMS	LHCb
Run 2	2 PB	0.5 PB	2 PB	10 PB (including Run 1)
Run 3	4 PB	1 PB	4 PB	45 PB
Total	6 PB	1.5 PB	6 PB	55 PB

<https://indico.cern.ch/event/1234612/contributions/5256185/>

Not scalable for Run 2 data

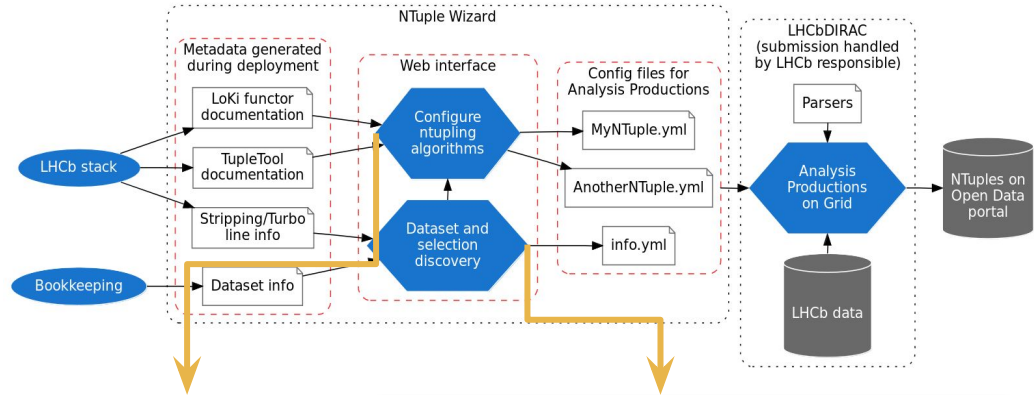
Run 2 Open Data Release

- ▶ Not feasible to repeat Run 1 release format for Run 2 data
- ▶ **Instead:** We release the data on Ntuple level upon Open Data requests
- ▶ **Ntuple Wizard** generates configuration files based on the user request [CHEP2023](#) [arXiv:2302.14245 \[hep-ex\]](#)
- ▶ Interpreted by LHCb internal production system (Analysis Productions) [CHEP2023](#) [Nicole's talk](#)
- ▶ **Ntupling Service** as gateway between **Ntuple Wizard**, Analysis Productions and Open Data Portal

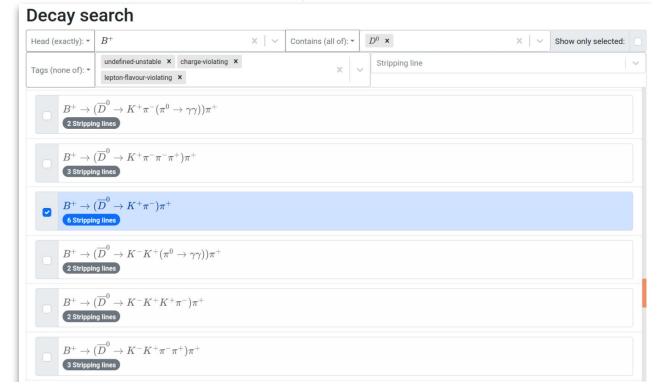
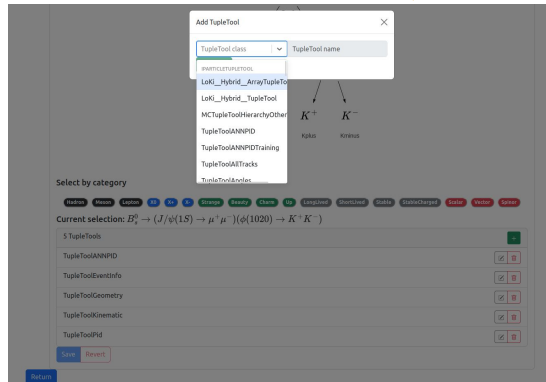


LHCb Ntuple Wizard

- ▶ LHCb dataset metadata is generated in the backend
- ▶ Frontend displays metadata in userfriendly interface
- ▶ Generates configuration files that can be parsed by internal production system!

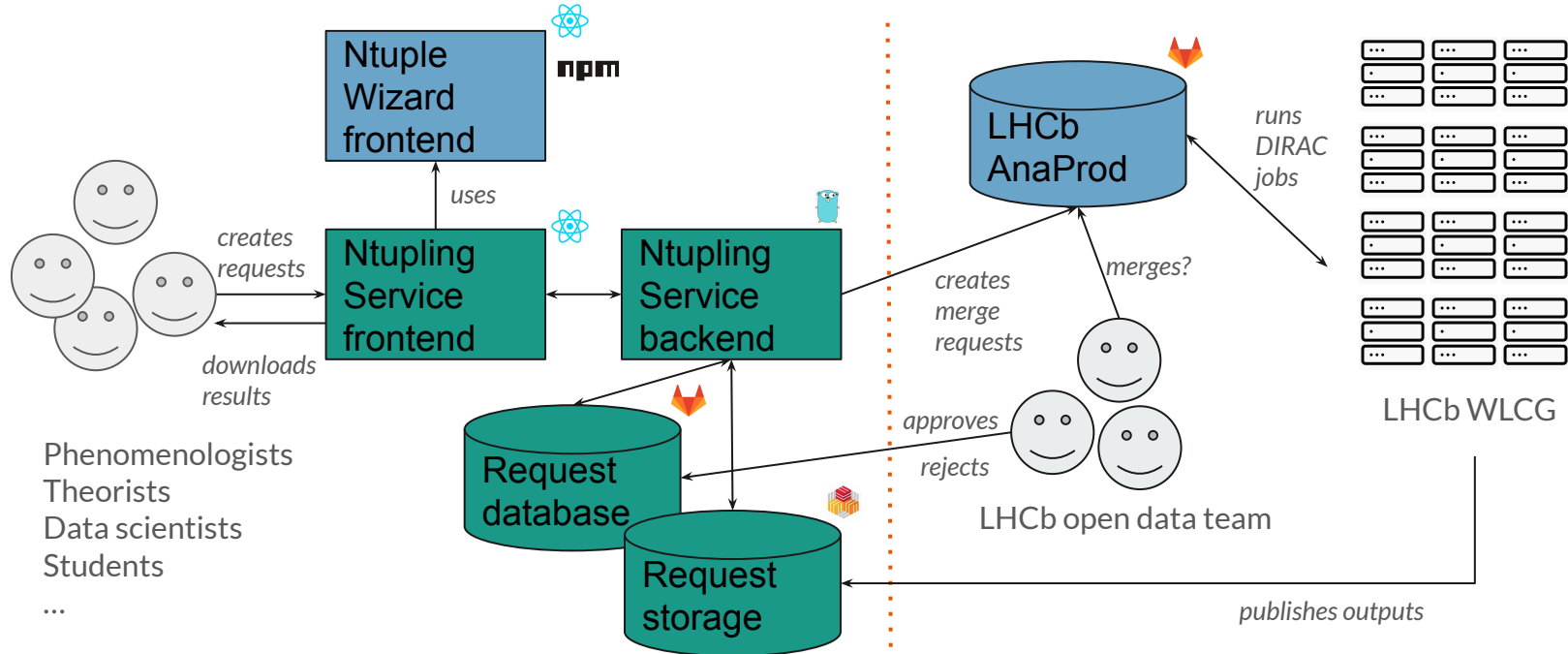


The **LHCb Ntupling Service** allows the release of the **Ntuple Wizard** as large scale Open Data project hosted on the Open Data Portal!



LHCb Ntupling Service

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

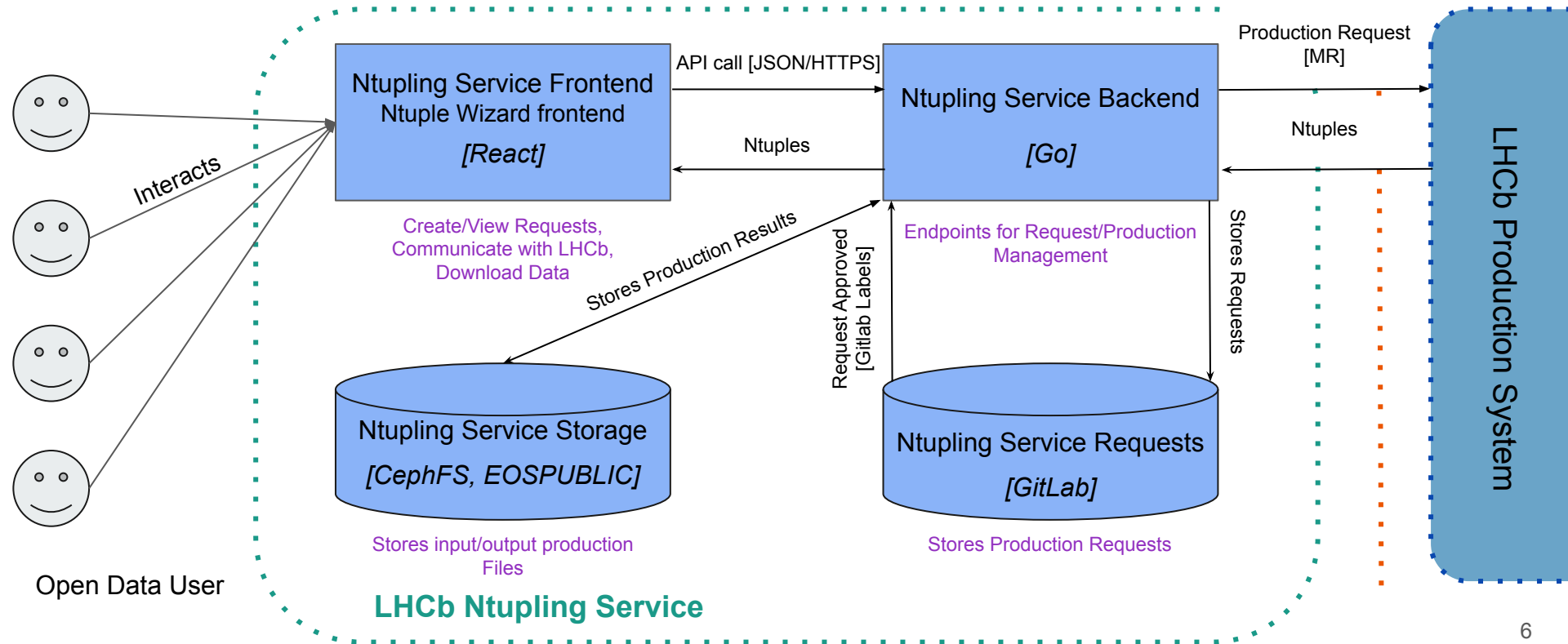


LHCb Ntupling Service Workflow

IT

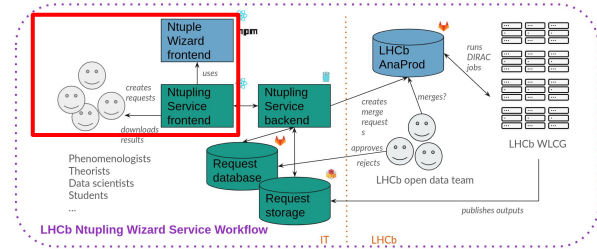
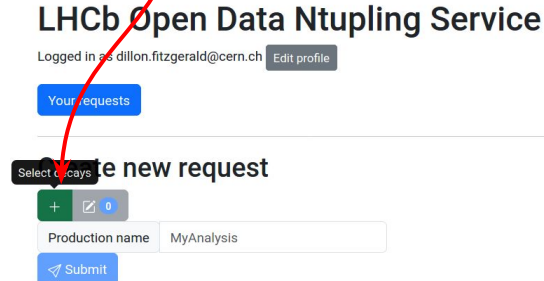
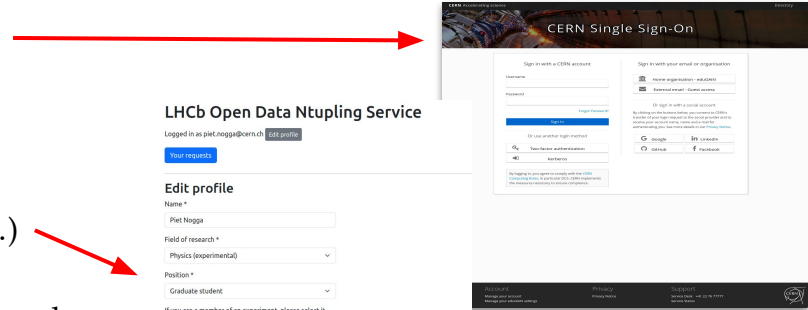
LHCb

LHCb Ntupling Service - Infrastructure



LHCb Open Data Ntupling Service - Step I

- ▶ User identification required for safety purposes
- ▶ Includes guest access or social login (Google, GitHub, ...)
- ▶ LHCb will ask to create a user profile (field of research, ...)
- ▶ Ntuple Wizard is integrated in the Ntupling Service Frontend



LHCb Open Data Ntupling Service - Step II

- Choose your data according to your desired decay!

Your requests Create new request

Create new request

DecayTree

$B_s^0 \rightarrow (J/\psi(1S) \rightarrow \mu^+ \mu^-)(\phi(1020) \rightarrow K^+ K^-)$

+ Select decays

Production name

StrippingBs2MuLinesBs2JPsiP...

LEPTONIC.MDST

LEPTONIC.MDST

Head (exactly): B_s^0 Contains (all of): $J/\psi(1S)$ $\phi(1020)$ Show only selected:

Tags (none of): undefined-unstable charge-violating Stripping line

11 Stripping lines

$B_s^0 \rightarrow (J/\psi(1S) \rightarrow \mu^+ e^-)(\phi(1020) \rightarrow K^+ K^-)$
1 Stripping line **lepton-flavour-violating**

$B_s^0 \rightarrow (J/\psi(1S) \rightarrow \mu^+ \mu^+ \mu^-)(\phi(1020) \rightarrow K^+ K^-)$
3 Stripping lines

$B_s^0 \rightarrow (J/\psi(1S) \rightarrow \mu^+ \mu^-)(\phi(1020) \rightarrow K^+ K^-)$
6 Stripping lines

$B_s^0 \rightarrow (J/\psi(1S) \rightarrow \mu^- e^+)(\phi(1020) \rightarrow K^+ K^-)$
1 Stripping line **lepton-flavour-violating**

$B_s^0 \rightarrow (J/\psi(1S) \rightarrow \mu^+ \mu^-)(\phi(1020) \rightarrow K^+ K^-) \mu^+ \mu^-$
1 Stripping line

$B_s^0 \rightarrow (\psi(2S) \rightarrow (J/\psi(1S) \rightarrow \mu^+ \mu^-) \pi^+ \pi^-)(\phi(1020) \rightarrow K^+ K^-)$
1 Stripping line

- The tuples are fully customizable
- *TupleTools* can be added per node of the interactive tree
- Writes collection of variables to Ntuple

[Abhijit's talk](#)

Add TupleTool

TupleTool class TupleTool name

LoKi_Hybrid_Array/TupleTool
LoKi_Hybrid_TupleTool
MCTupleToolHierarchyOther
TupleToolANNPD
TupleToolANNPDTraining
TupleToolAllTracks
TupleToolAxes

Select by category

Your requests Create new request

Create new request

DecayTreeTuple configuration

Tree name: DecayTree

1 input location

Configure $B_s^0 \rightarrow (J/\psi(1S) \rightarrow \mu^+ \mu^-)(\phi(1020) \rightarrow K^+ K^-)$

B_s^0

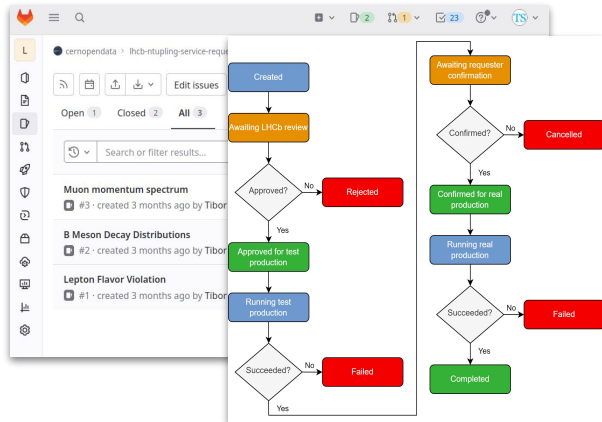
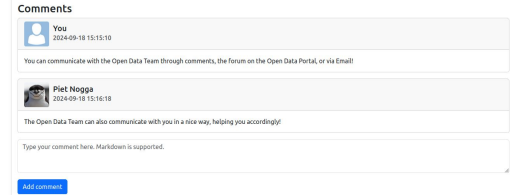
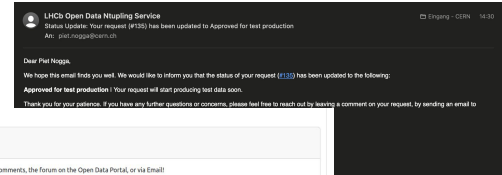
$J/\psi(1S)$ $\phi(1020)$

μ^+ μ^- K^+ K^-

muplus minus Kplus Kminus

LHCb Open Data Ntupling Service - Step III

- ▶ Once satisfied, you can submit your request and it is picked up by the **Ntupling Service**
- ▶ It is always possible to view and edit your requests via the Open Data Portal
- ▶ Communicate with the Open Data Team at any point during the request!



- ▶ Service allocates unique request ID and saves configuration files internally
- ▶ Open Data Team can discuss the requests internally via GitLab issues
- ▶ Requests are individually reviewed by the Open Data Team

LHCb Open Data Ntupling Service - Step IV

- ▶ As soon as the request is approved, the **Ntupling Service** will automatically open a MR on Analysis Productions
- ▶ This triggers a small test production, from which the request's relevant information is extracted
- ▶ The information is written to a markdown file and automatically propagated to Service Frontend

The test production for your request has been completed! Please review the results to ensure that the outcome matches your expectations, and if so, approve the request for actual real production.

Test production results

Data	Size		
test-production.md	20.876 KiB		

I confirm that the test production gave the expected results and I would like to execute the real production.

Submit

Ntupling Service Request | test_examples | dillon.fitzgerald@cern.ch

[View](#) [LHCb Open Data Requestor](#) requested to merge [dillonf-test-tup-service...](#) [Info](#) [History](#) 1 week ago

Overview 2 | Commits 3 | Pipelines 3 | Changes 7

Requester

- Name: Dillon Fitzgerald
- Email: dillon.fitzgerald@cern.ch
- Field of research: Physics (experimental)
- Position: Graduate student
- Experiment: LHCb
- Remarks: Testing the system

Request Details

- Production name: test_examples

B2JpsiKK will process approximately 53849.5 GB of data and is expected to generate around 29.3 GB of output. Please note that the actual size of the output files may vary from the estimate.

- ▶ **See Branches in Production Tuples**

N.B. this is an automated message from the friendly Analysis Productions Bot.

- ▶ You can take a look at the produced variables in the final tuple and its estimated output size
- ▶ Confirm test result to trigger the full production of your tuple!

Labels Edit

Running test production ×

Running test production

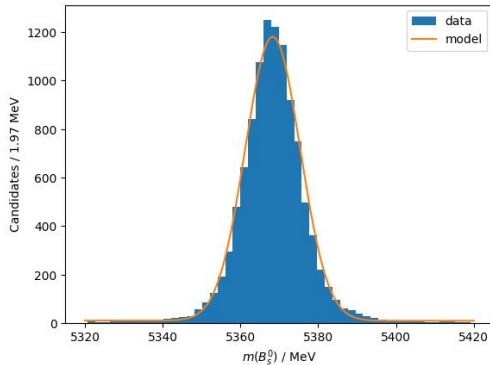
Select labels ×

Search labels

- Approved for test production
- Awaiting LHCb review
- Awaiting requester confirmation

LHCb Open Data Ntupling Service - Step V

- ▶ LHCb Open Data Team merges the request and the production job runs on the grid
- ▶ Output ROOT files are transferred to *eospublic*, and propagated to the Frontend



- ▶ Service will allow promoting Ntuples as Open Data records with DOI
- ▶ You can now start your own analysis using LHCb data!

Labels
Completed ✕

We are processing 5.35 Million Candidates!

Candidates / #00

10000
80000
60000
40000
20000
0

5150 5200 5250 5300 5350 5400 5450 5500 5550

m[lpK*K*]

open data
cm

Important notice: This application is un...

LHCb Open Data Ntupli

Logged in as tiborsimko@cern.ch Edit profile

Your request: [Completed]

Your ntuples

For demonstration purposes only
This page has been created solely for demonstration purposes and cannot be used to view and download your ntuples as of yet.

All ntuples are removed from the system after 90 days. If you want to keep your ntuples, please download them in time.

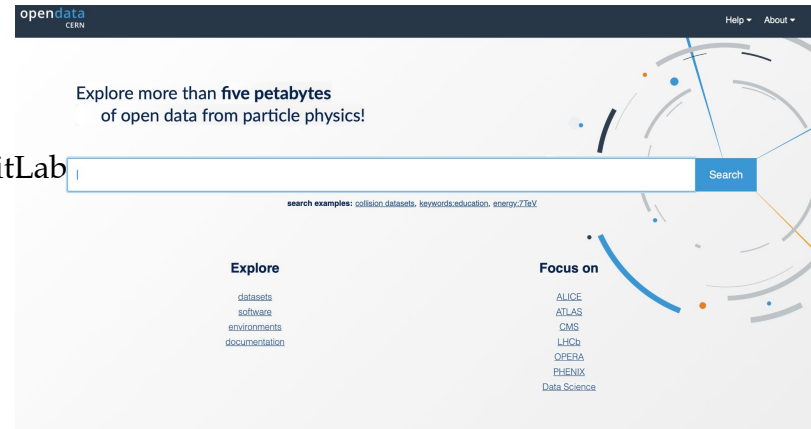
Filename	Expires in	Size	Produced by			
B2HHH_MagnetDown.root	10 days	636 MiB	B Meson Decay Distributions	1	2	3
PhaseSpaceSimulation.root	10 days	2 MiB	B Meson Decay Distributions	1	2	3
B2psiKSPITuple_2011MagDown.root	10 days	5 MiB	B Meson Decay Distributions	1	2	3
00012345_00006789_1_dtuple.root	10 days	88 KiB	B Meson Decay Distributions	1	2	3

We hope this email finds you well. We would like to inform you that the status of your request ([#137](#)) has been updated to the following:

Completed | Your request has generated the requested data. You can download the results now.

Summary

- ▶ Open Data releases are essential for promoting **transparency and collaborative research**
- ▶ The LHCb Collaboration needs a new approach for Run 2 Open Data release compared to the Run 1 release
- ▶ The **LHCb Open Data Ntupling Service** is being developed in a close collaboration between the CERN IT Department and the LHCb Open Data Team
- ▶ Users interact with the Service Frontend, embedding the Ntuple Wizard
- ▶ Requests are handled in the Service Backend, storing them in a GitLab database and propagating configuration files to LHCb Production System
- ▶ **Novel approach to publish open data via runnable workflows**



Outlook

- ▶ The application is in an LHCb internal *alpha* phase since February 2024
- ➔ Feedback from LHCb Collaboration
- ▶ LHCb Ntupling Service is currently being presented at the LHCb implications workshop
- ➔ Entering *beta* phase and implementing feedback from affiliated theorists



**First public release of LHCb Ntupling Service
expected in 2025, efforts ongoing to release all LHCb**

Run 2 data by 2028!

Stay tuned!

