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

Christine Aidala[1], Dillon Fitzgerald[1], Kai Habermann[2], Ludwig Kramer[2], Adam Morris,[3] Sebastian Neubert[2], **Piet Nogga**[2], Eduardo Rodrigues[4], Marco Donadoni[3], Daan Rosendal[3], Tibor Simko[3]

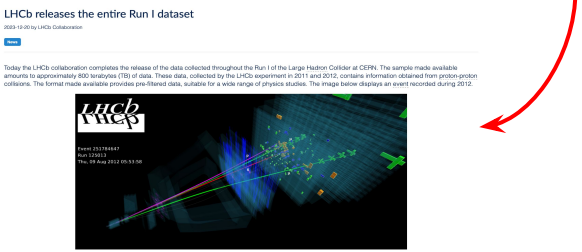
[1]University of Michigan, [2]University of Bonn,
[3]CERN, [4]Univerity of Liverpool

Bundesministerium für Bildung und Forschung

October 19 - 25, 2024
CHEP 2024

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!



https://opendata.cern.ch

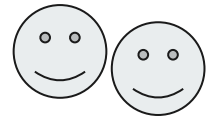
	ALICE	ATLAS	CMS	LHCb
Run 2	2 PB	0.5 PB	2 PB	10 PB (including Run 1)
Run 3	4 PB	1 PB	4 PB	45 PB
Total	**6 PB**	**1.5 PB**	**6 PB**	**55 PB**

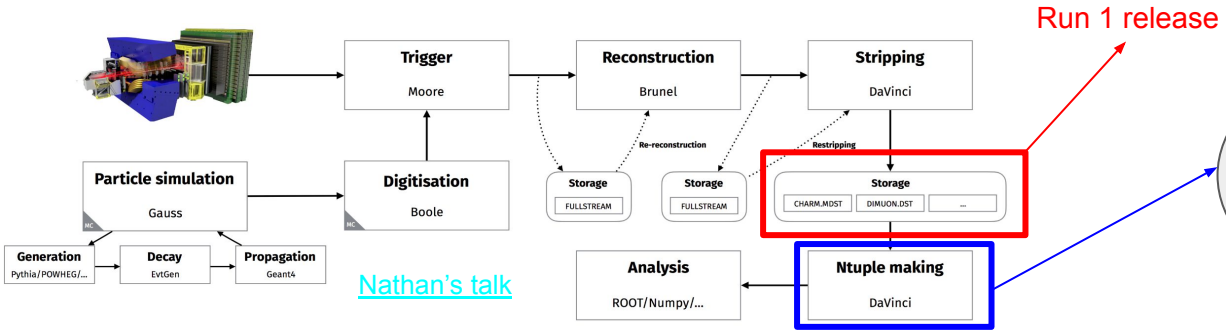
https://indico.cern.ch/event/1234612/contributions/5253185/

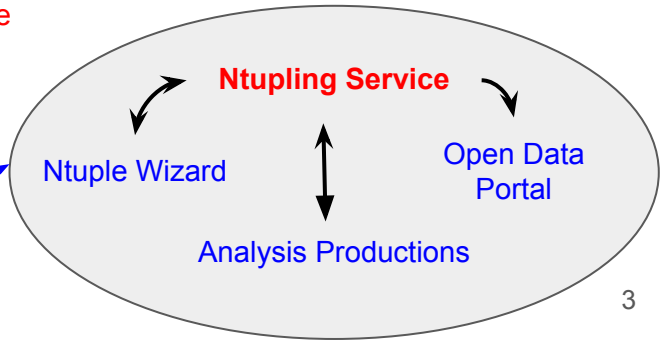
Not scalable for Run 2 data

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

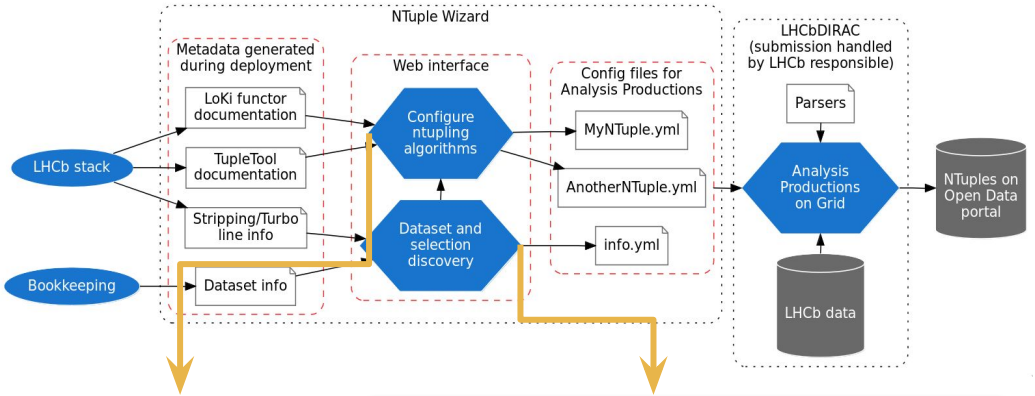


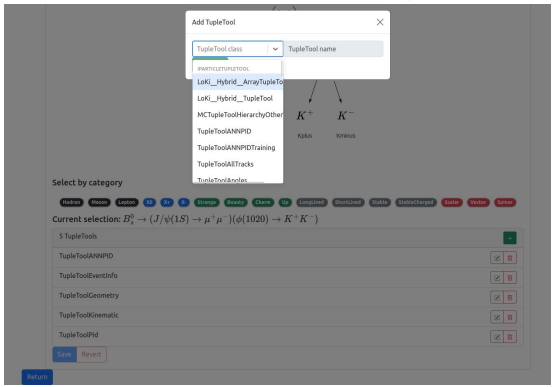


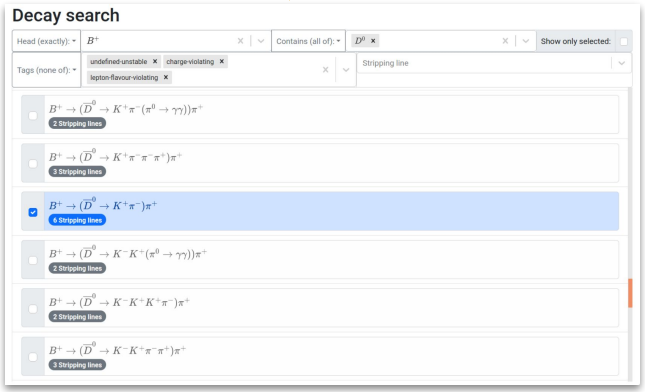


LHCb Ntuple Wizard

- 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!



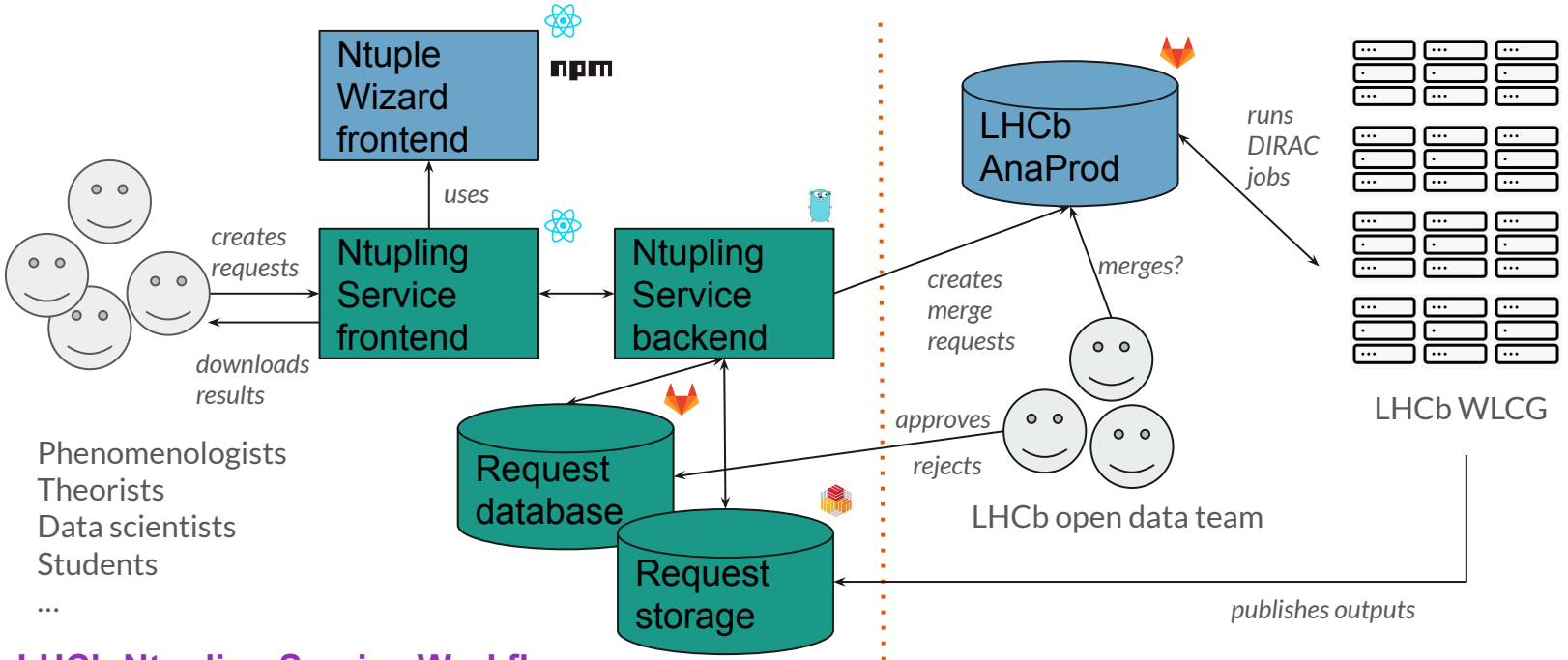




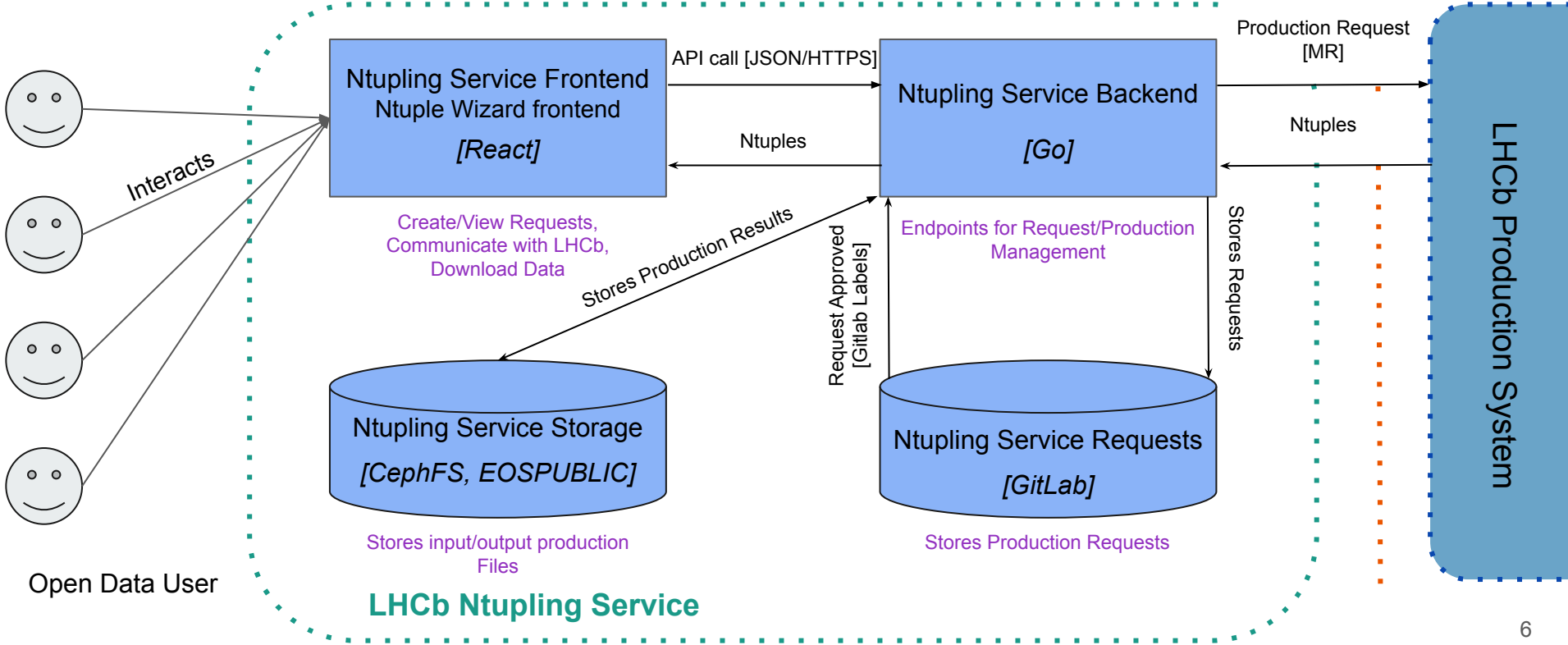
The **LHCb Ntupling Service** allows the release of the Ntuple Wizard as large scale Open Data project hosted on the Open Data Portal!

LHCb Ntupling Service

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

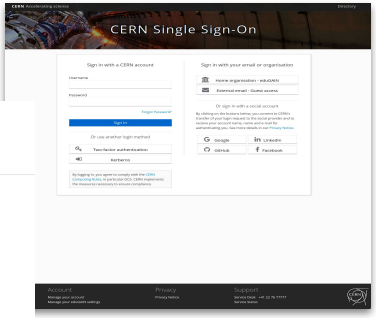


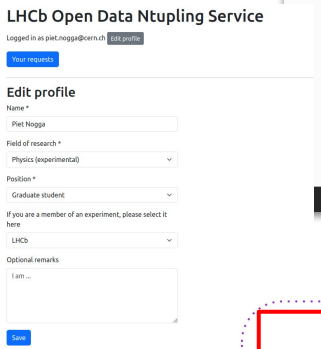
LHCb Ntupling Service - Infrastructure

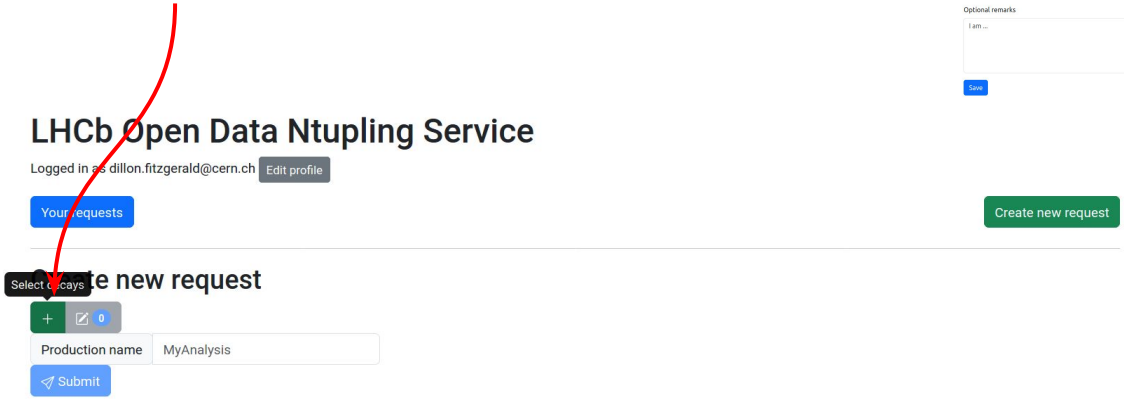


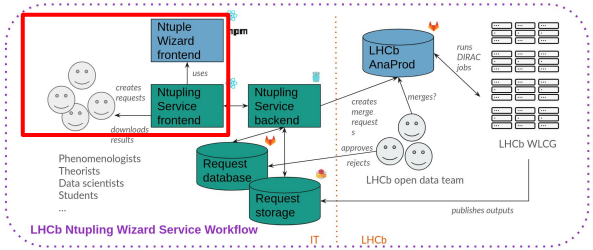
LHCb Open Data Ntupling Service - Step I

- 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







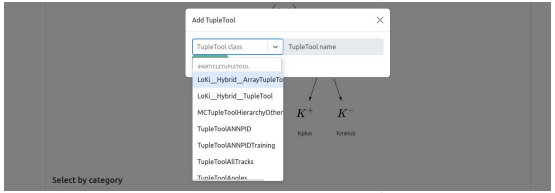


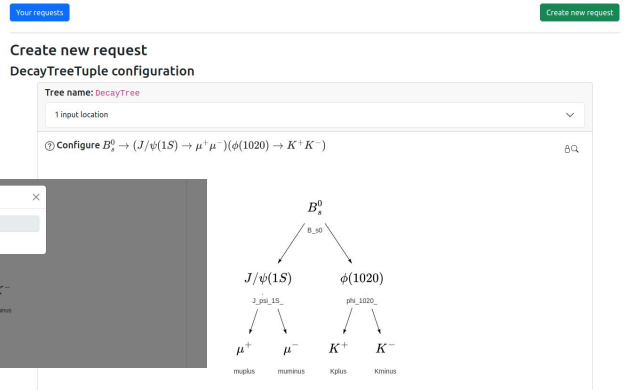
LHCb Open Data Ntupling Service - Step II

- Choose your data according to your desired decay!

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

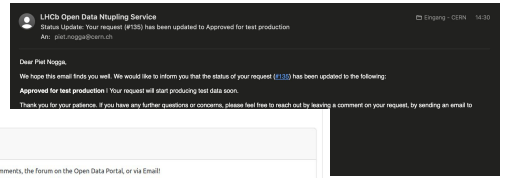
Abhijit's talk

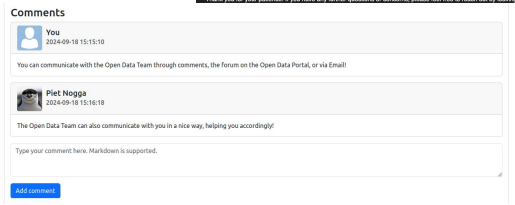


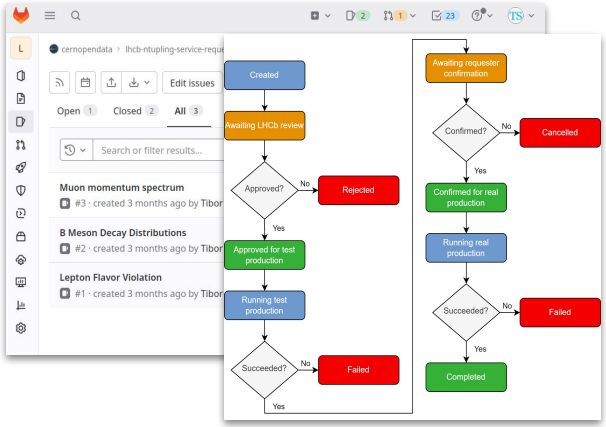


LHCb Open Data Ntupling Service - Step III

- 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!



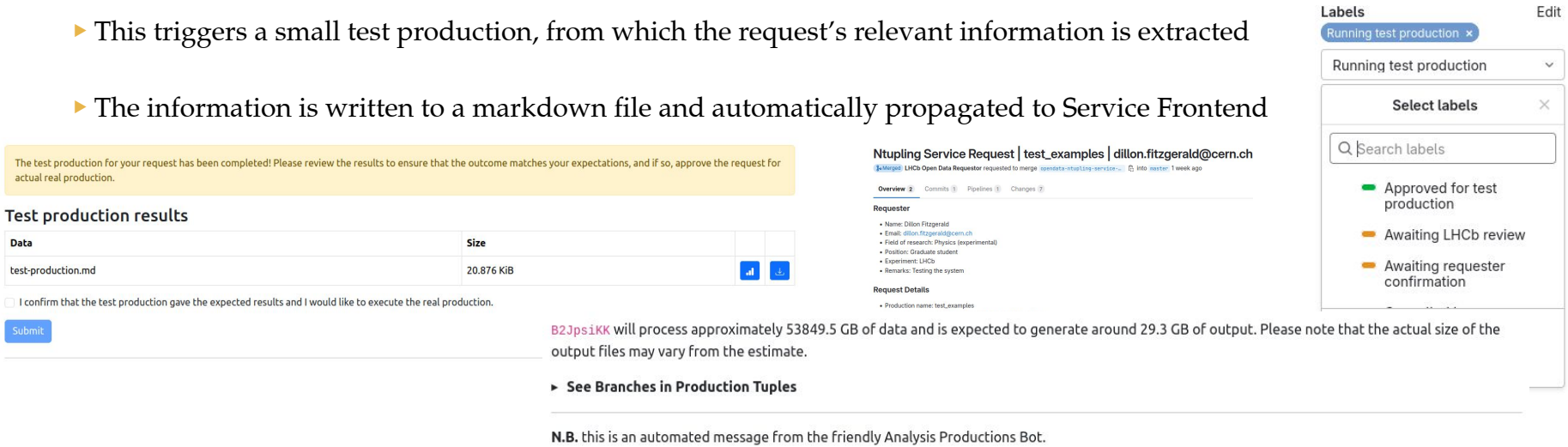




- Service allocates unique request ID and saves configuration files internally
- Open Data Team can discuss the requests internally via GitLab issues
- Requests are individually reviewed by the Open Data Team

LHCb Open Data Ntupling Service - Step IV

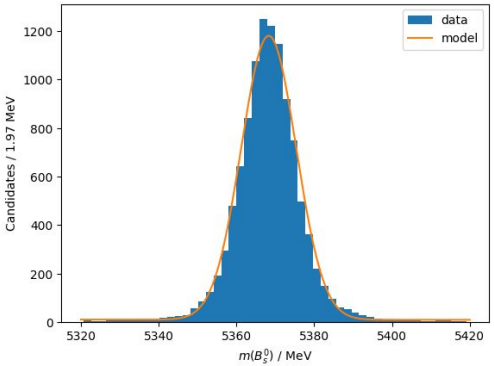
- As soon as the request is approved, the Ntupling Service will automatically open a MR on Analysis Productions
- This triggers a small test production, from which the request's relevant information is extracted
- The information is written to a markdown file and automatically propagated to Service Frontend



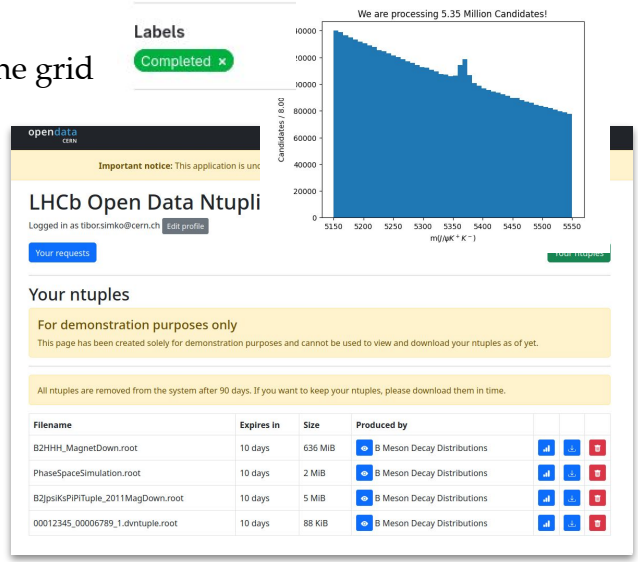
- You can take a look at the produced variables in the final tuple and its estimated output size
- Confirm test result to trigger the full production of your tuple!

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!

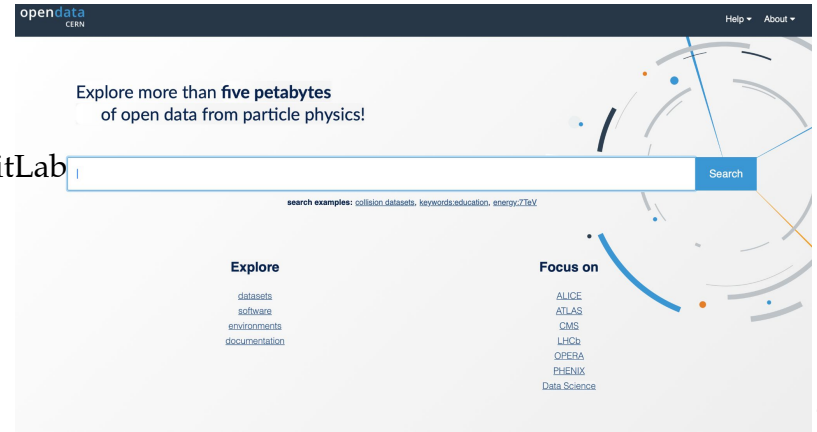


We hope this email finds you well. We would like to inform you that the status of your request (#137) has been updated to the following:

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!