Norfolk, Virginia, USA • May 8-12, 2023
CHEP 2023
Computing in High Energy & Nuclear Physics

An Ntuple production service for accessing LHCb Open Data: the Ntuple Wizard

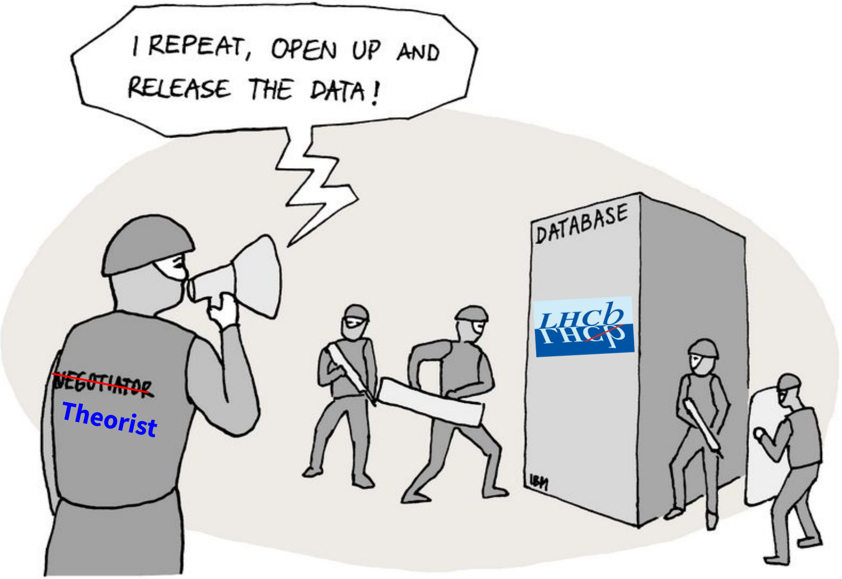
Dillon S. Fitzgerald on behalf of the LHCb collaboration

arXiv:2302.14235 [hep-ex]

May 9, 2023

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!)



Dataedo /cartoon

Piotr@Dataedo

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

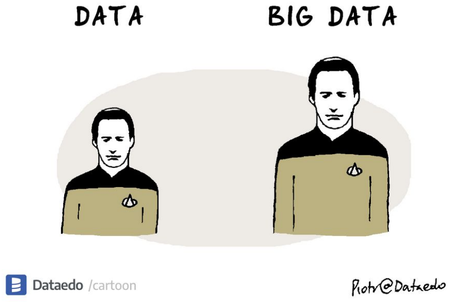
- Level 4: Raw data
  - the basic quantities recorded by the experimental instruments.

LHCb Open Data

LHCb recently released about 20% the Run 1 data (200 TB) on the CERN Open Data Portal: https://opendata.cern.ch/search?page=1&size=20&experiment=LHCb

Additional releases 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**



Note: Flavour physics analyses often require much more event and decay information compared to typical analyses on other LHC experiments

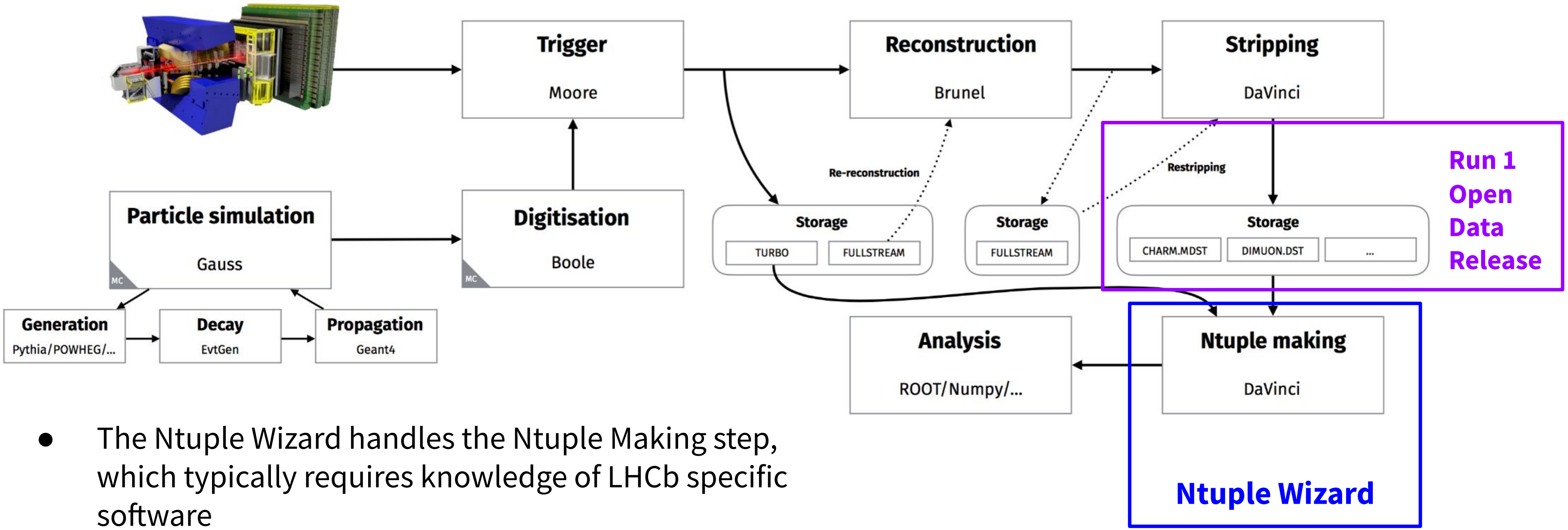
This is not scalable! This prompted the development of a new system…

The Ntuple Wizard

LHCb Run 1 and 2 Data Flow

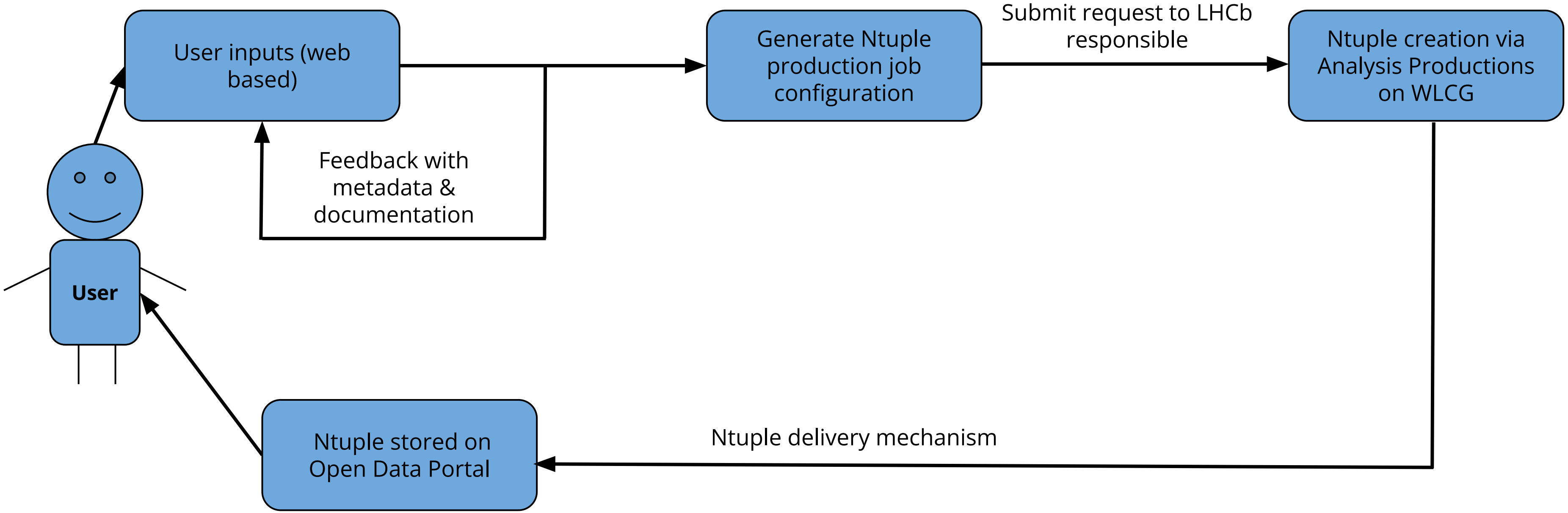
Stripping = skimming + trimming

Reconstructed events are filtered to create collections with particular physics signatures

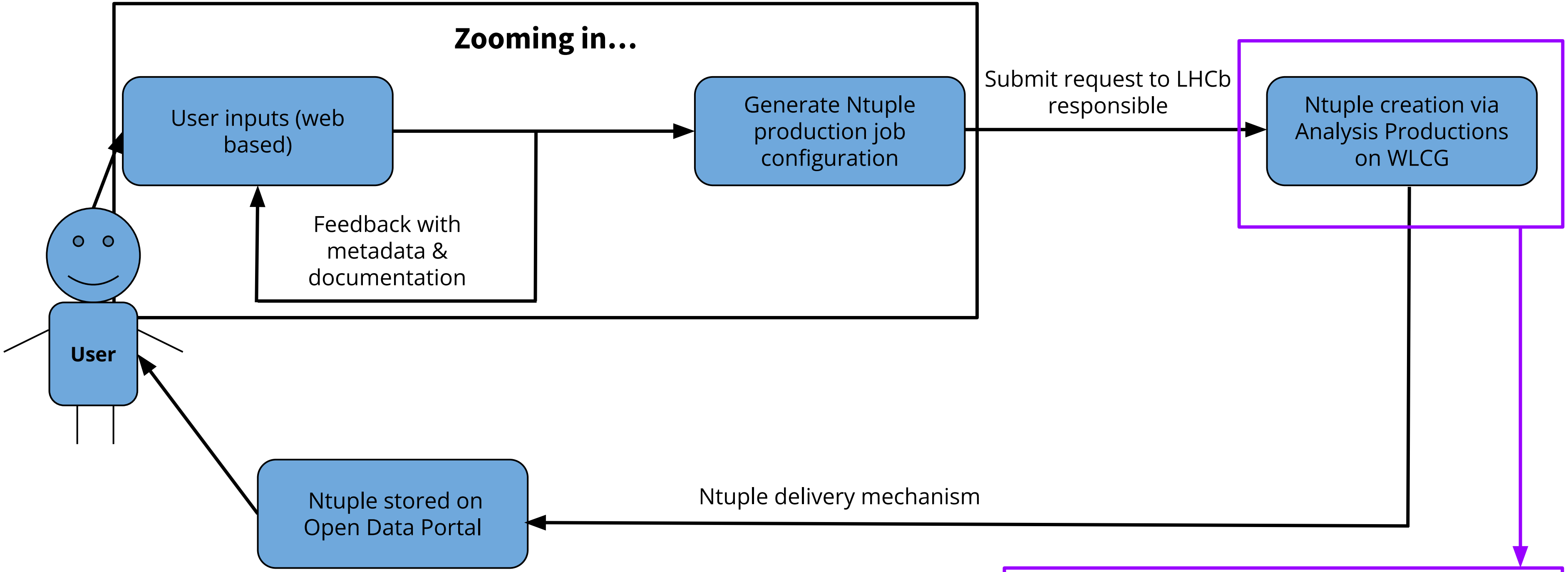


- The Ntuple Wizard handles the Ntuple Making step, which typically requires knowledge of LHCb specific software
    - **Lower barrier of entry for external analysts!**

The Ntuple Wizard


User inputs (web based)
Generate Ntuple production job configuration
Submit request to LHCb responsible
Ntuple creation via Analysis Productions on WLCG
Feedback with metadata & documentation
User
Ntuple delivery mechanism
Ntuple stored on Open Data Portal

The Ntuple Wizard



Zooming in…

User inputs (web based)

Generate Ntuple production job configuration

Feedback with metadata & documentation

Submit request to LHCb responsible

Ntuple creation via Analysis Productions on WLCG

User

Ntuple delivery mechanism

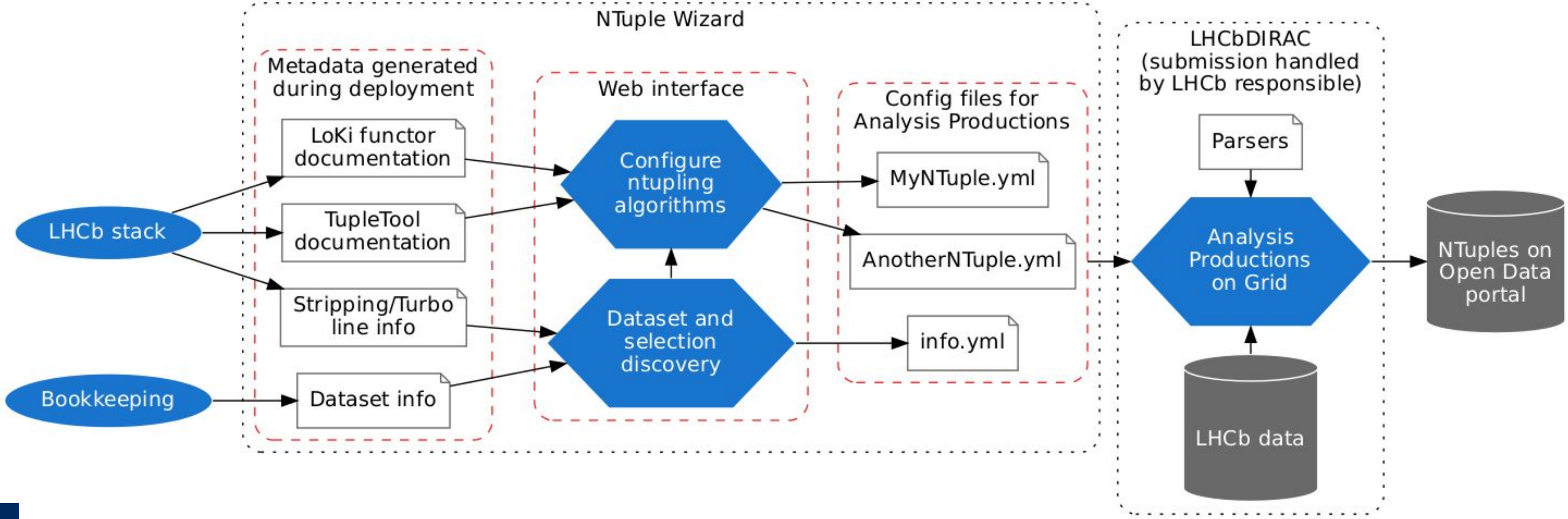
Ntuple stored on Open Data Portal

See detailed talk on Analysis Productions by Chris Burr -- May 9 @ 5:45 PM

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

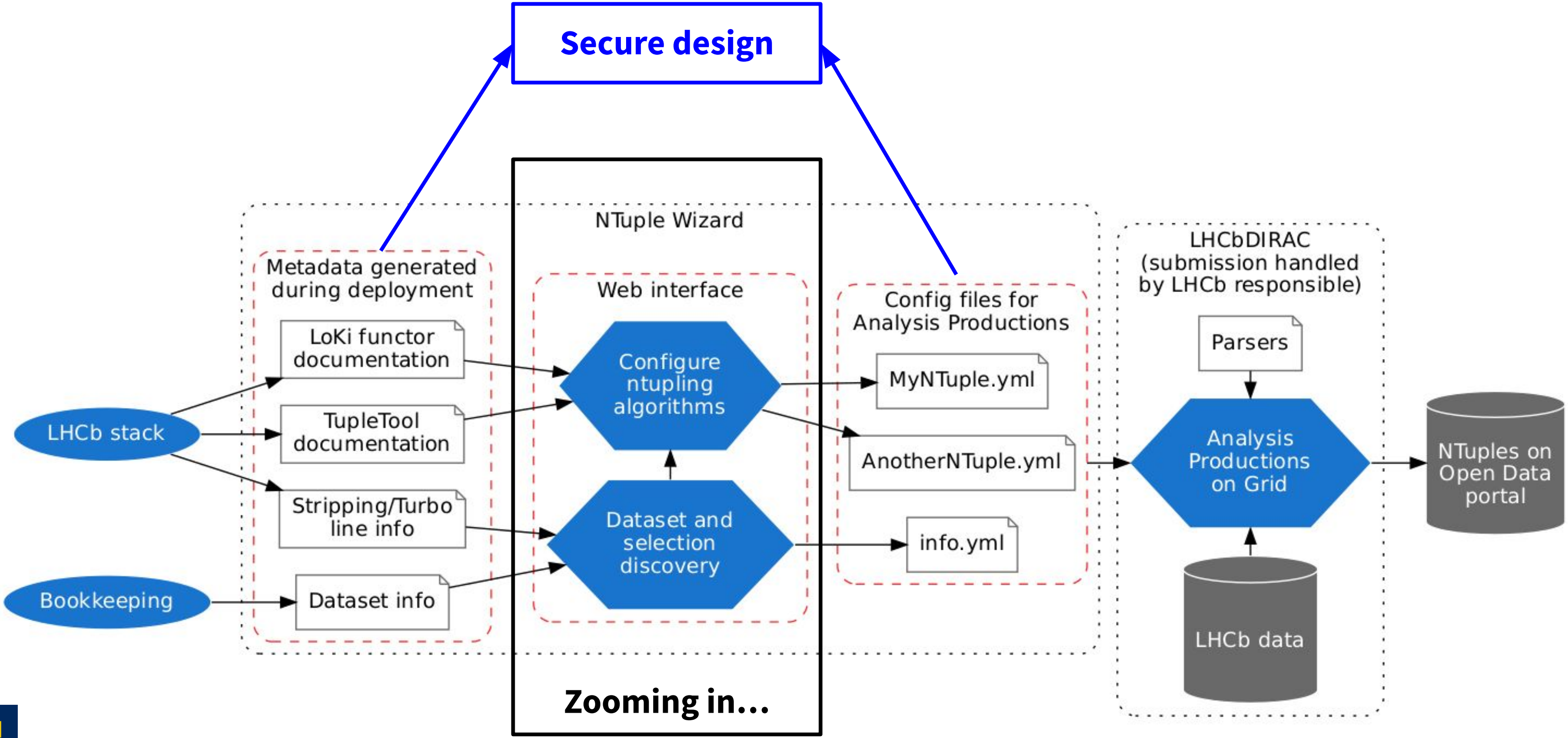


NTuple Wizard
Metadata generated during deployment
Web interface
Config files for Analysis Productions
LHCbDIRAC (submission handled by LHCb responsible)
LHCb stack
Bookkeeping
LoKi functor documentation
TupleTool documentation
Stripping/Turbo line info
Dataset info
Configure ntupling algorithms
Dataset and selection discovery
MyNTuple.yml
AnotherNTuple.yml
info.yml
Parsers
Analysis Productions on Grid
LHCb data
NTuples on Open Data portal

Interlude: Security & Permissions

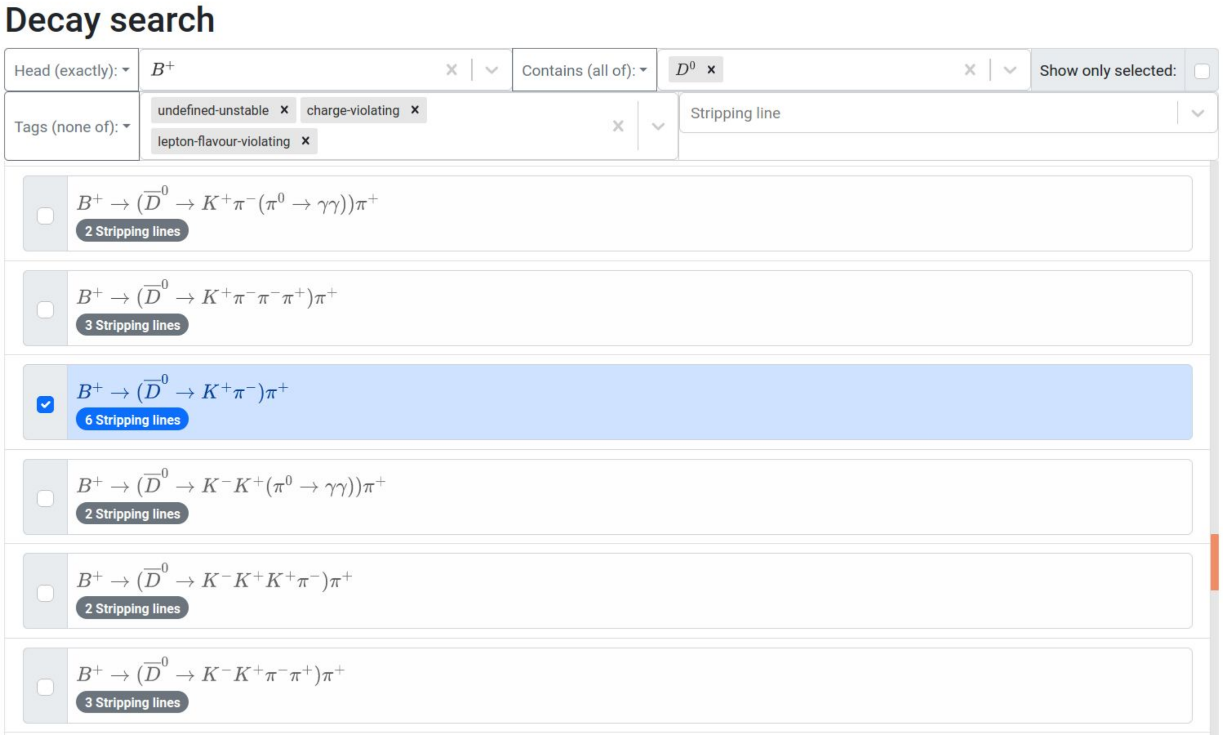
- Standard LHCb Ntuple making application (DaVinci) configured with python scripts
  - **Running arbitrary code from external users is a security risk!**
  - **Config output saved in yaml data structures, interpreted by internal parsers**
- Dataset discovery and Ntuple configuration require metadata 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 only contains a subset of the data because of this
  - The Ntuple Wizard can improve this via **fine-grained control** over:
    - building/accessing decay candidates
    - Stripping lines or equivalent selections

The Ntuple Wizard


Secure design
NTuple Wizard
Metadata generated during deployment
LoKi functor documentation
TupleTool documentation
Stripping/Turbo line info
Dataset info
LHCb stack
Bookkeeping
Web interface
Configure ntupling algorithms
Dataset and selection discovery
Config files for Analysis Productions
MyNTuple.yml
AnotherNTuple.yml
info.yml
LHCbDIRAC (submission handled by LHCb responsible)
Parsers
Analysis Productions on Grid
LHCb data
NTuples on Open Data portal
Zooming in…

Web Interface: Dataset Discovery

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



Decay search

Head (exactly): B^+ | Contains (all of): D^0 | Show only selected: □

Tags (none of): undefined-unstable ✕ charge-violating ✕ lepton-flavour-violating ✕ | Stripping line

□ $B^+ \to (\overline{D}^0 \to K^+\pi^-(\pi^0 \to \gamma\gamma))\pi^+$ 2 Stripping lines

□ $B^+ \to (\overline{D}^0 \to K^+\pi^-\pi^-\pi^+)\pi^+$ 3 Stripping lines

☑ $B^+ \to (\overline{D}^0 \to K^+\pi^-)\pi^+$ 6 Stripping lines

□ $B^+ \to (\overline{D}^0 \to K^-K^+(\pi^0 \to \gamma\gamma))\pi^+$ 2 Stripping lines

□ $B^+ \to (\overline{D}^0 \to K^-K^+K^+\pi^-)\pi^+$ 2 Stripping lines

□ $B^+ \to (\overline{D}^0 \to K^-K^+\pi^-\pi^+)\pi^+$ 3 Stripping lines

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

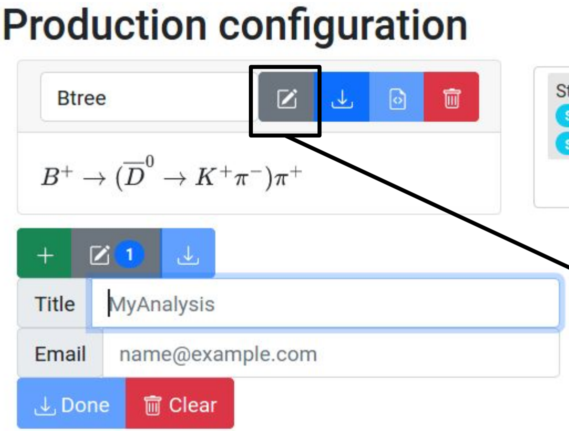
Selection of a physics object exposes the corresponding available datasets for the user to choose from

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



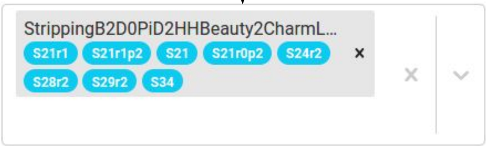
Production configuration

Btree

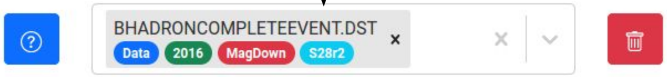
$B^+ \to (\overline{D}^0 \to K^+\pi^-)\pi^+$

Title: MyAnalysis

Email: name@example.com



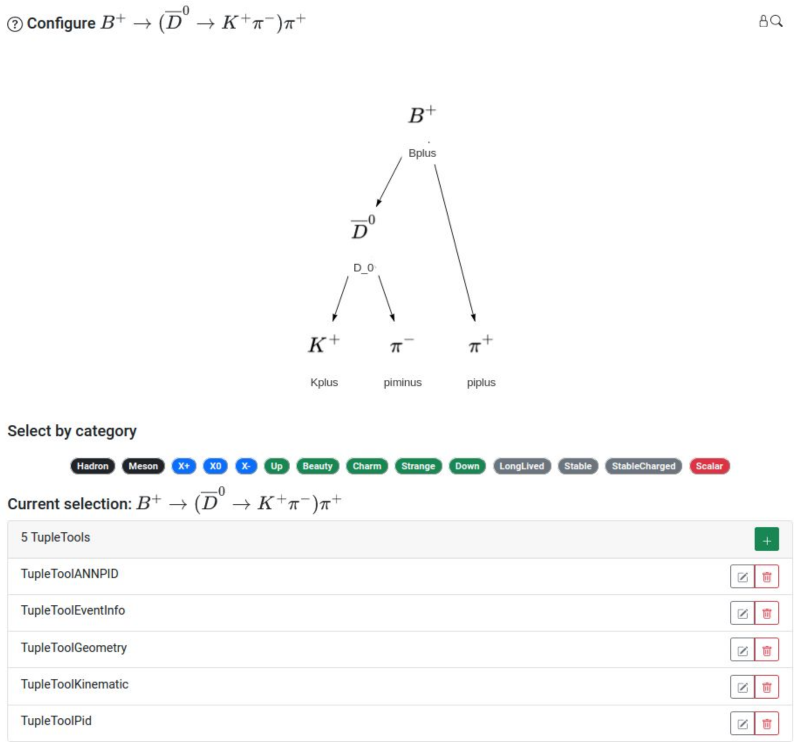
StrippingB2D0PiD2HHBeauty2CharmL...



BHADRONCOMPLETEEVENT.DST

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

Web Interface: Ntuple Configuration



Configure $B^+ \to (\overline{D}^0 \to K^+\pi^-)\pi^+$

Select by category

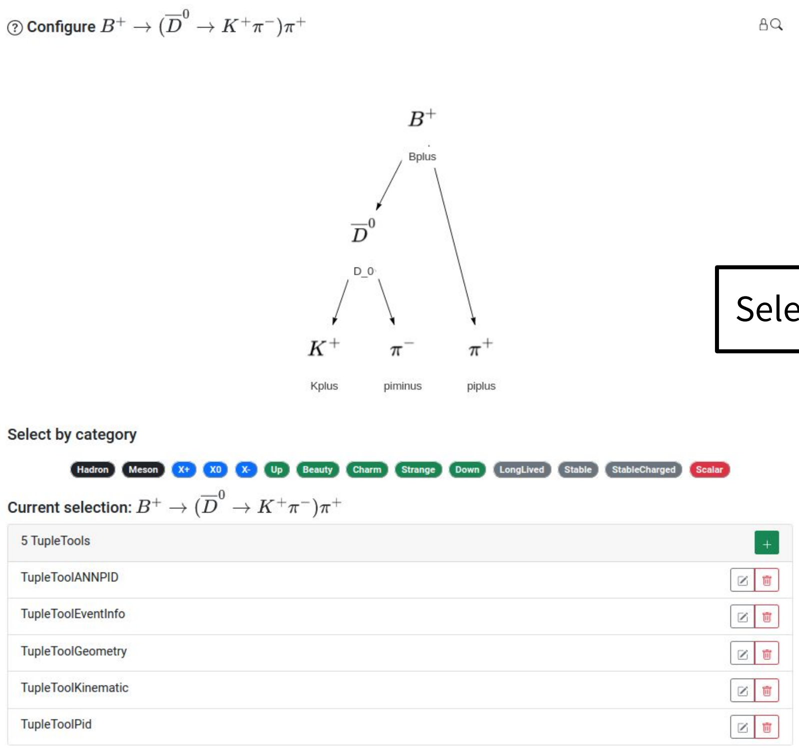
Current selection: $B^+ \to (\overline{D}^0 \to K^+\pi^-)\pi^+$

5 TupleTools
TupleToolANNPID
TupleToolEventInfo
TupleToolGeometry
TupleToolKinematic
TupleToolPid

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



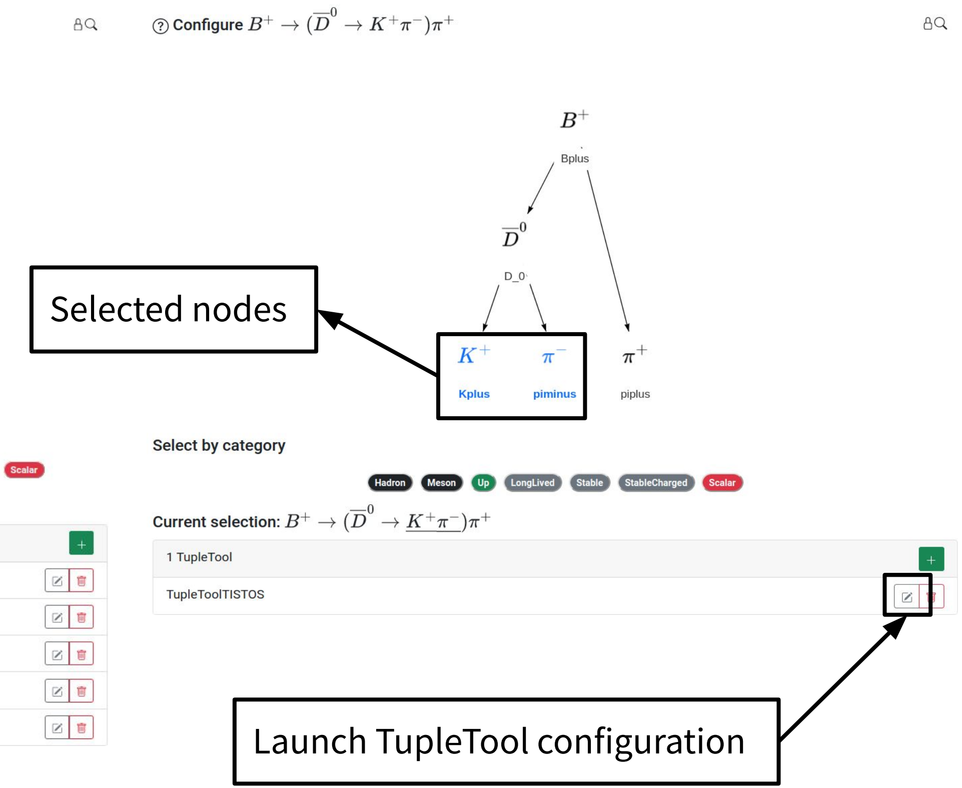
? Configure $B^+ \to (\overline{D}^0 \to K^+\pi^-)\pi^+$

Select by category

Current selection: $B^+ \to (\overline{D}^0 \to K^+\pi^-)\pi^+$

5 TupleTools +

- TupleToolANNPID
- TupleToolEventInfo
- TupleToolGeometry
- TupleToolKinematic
- TupleToolPid



? Configure $B^+ \to (\overline{D}^0 \to K^+\pi^-)\pi^+$

Select by category

Current selection: $B^+ \to (\overline{D}^0 \to \underline{K^+\pi^-})\pi^+$

1 TupleTool +

- TupleToolTISTOS

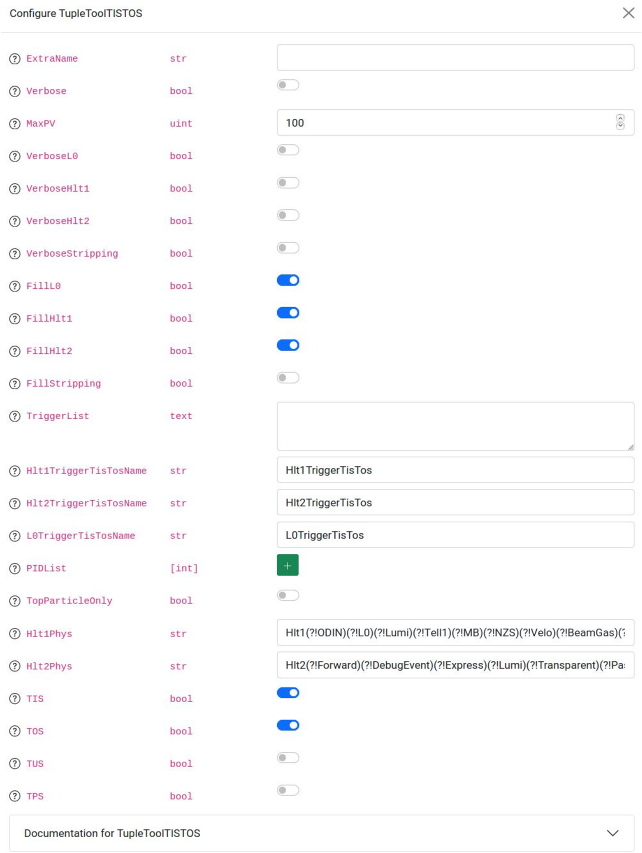
Selected nodes

Launch TupleTool configuration

Web Interface: Tuple Tool Configuration Example

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

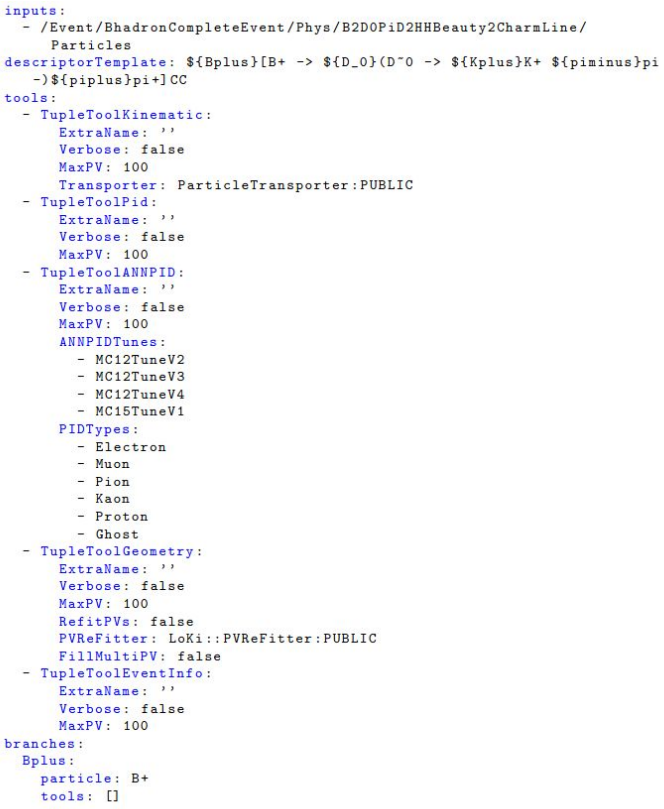


Configure TupleToolTISTOS

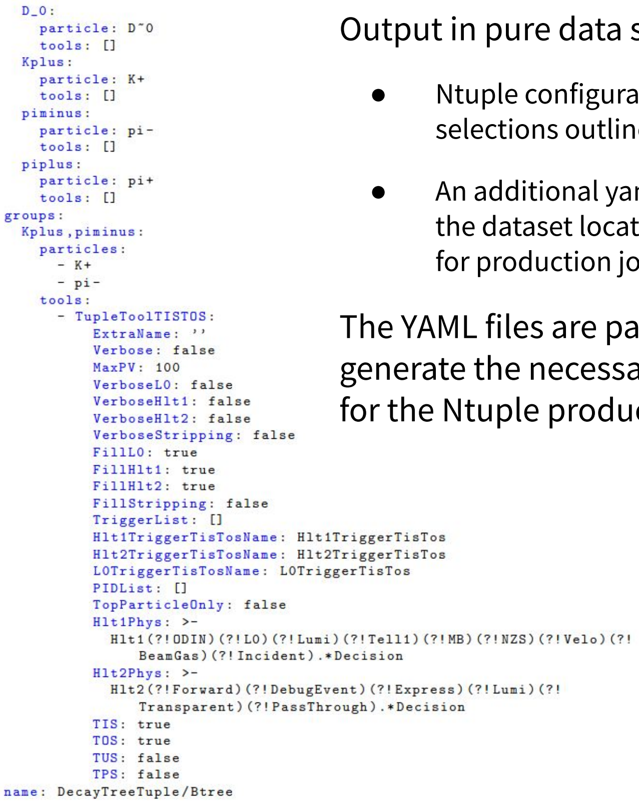
ExtraName	str	
Verbose	bool	○
MaxPV	uint	100
VerboseL0	bool	○
VerboseHlt1	bool	○
VerboseHlt2	bool	○
VerboseStripping	bool	○
FillL0	bool	●
FillHlt1	bool	●
FillHlt2	bool	●
FillStripping	bool	○
TriggerList	text	
Hlt1TriggerTisTosName	str	Hlt1TriggerTisTos
Hlt2TriggerTisTosName	str	Hlt2TriggerTisTos
L0TriggerTisTosName	str	L0TriggerTisTos
PIDList	[int]	+
TopParticleOnly	bool	○
Hlt1Phys	str	Hlt1(?!ODIN)(?!L0)(?!Lumi)(?!Tell1)(?!MB)(?!NZS)(?!Velo)(?!BeamGas)(?
Hlt2Phys	str	Hlt2(?!Forward)(?!DebugEvent)(?!Express)(?!Lumi)(?!Transparent)(?!Pa
TIS	bool	●
TOS	bool	●
TUS	bool	○
TPS	bool	○

Documentation for TupleToolTISTOS

Ntuple Configuration Output Example



```
inputs:
  - /Event/BhadronCompleteEvent/Phys/B2D0PiD2HHBeauty2CharmLine/
     Particles
descriptorTemplate: ${Bplus}[B+ -> ${D_0}(D~0 -> ${Kplus}K+ ${piminus}pi
   -)${piplus}pi+]CC
tools:
  - TupleToolKinematic:
      ExtraName: ''
      Verbose: false
      MaxPV: 100
      Transporter: ParticleTransporter:PUBLIC
  - TupleToolPid:
      ExtraName: ''
      Verbose: false
      MaxPV: 100
  - TupleToolANNPID:
      ExtraName: ''
      Verbose: false
      MaxPV: 100
      ANNPIDTunes:
        - MC12TuneV2
        - MC12TuneV3
        - MC12TuneV4
        - MC15TuneV1
      PIDTypes:
        - Electron
        - Muon
        - Pion
        - Kaon
        - Proton
        - Ghost
  - TupleToolGeometry:
      ExtraName: ''
      Verbose: false
      MaxPV: 100
      RefitPVs: false
      PVReFitter: LoKi::PVReFitter:PUBLIC
      FillMultiPV: false
  - TupleToolEventInfo:
      ExtraName: ''
      Verbose: false
      MaxPV: 100
branches:
  Bplus:
    particle: B+
    tools: []
```



```
  D_0:
    particle: D~0
    tools: []
  Kplus:
    particle: K+
    tools: []
  piminus:
    particle: pi-
    tools: []
  piplus:
    particle: pi+
    tools: []
groups:
  Kplus,piminus:
    particles:
      - K+
      - pi-
    tools:
      - TupleToolTISTOS:
          ExtraName: ''
          Verbose: false
          MaxPV: 100
          VerboseL0: false
          VerboseHlt1: false
          VerboseHlt2: false
          VerboseStripping: false
          FillL0: true
          FillHlt1: true
          FillHlt2: true
          FillStripping: false
          TriggerList: []
          Hlt1TriggerTisTosName: Hlt1TriggerTisTos
          Hlt2TriggerTisTosName: Hlt2TriggerTisTos
          L0TriggerTisTosName: L0TriggerTisTos
          PIDList: []
          TopParticleOnly: false
          Hlt1Phys: >-
            Hlt1(?!ODIN)(?!L0)(?!Lumi)(?!Tell1)(?!MB)(?!NZS)(?!Velo)(?!
               BeamGas)(?!Incident).*Decision
          Hlt2Phys: >-
            Hlt2(?!Forward)(?!DebugEvent)(?!Express)(?!Lumi)(?!
               Transparent)(?!PassThrough).*Decision
          TIS: true
          TOS: true
          TUS: false
          TPS: false
name: DecayTreeTuple/Btree
```

Output in pure data structure (YAML) format

- Ntuple configuration output shown based on selections outlined in the previous slides
- An additional yaml file is generated to specify the dataset location and organize the request for production jobs (not shown here)

The YAML files are parsed internally to generate the necessary python options files for the Ntuple production jobs

Summary (1)

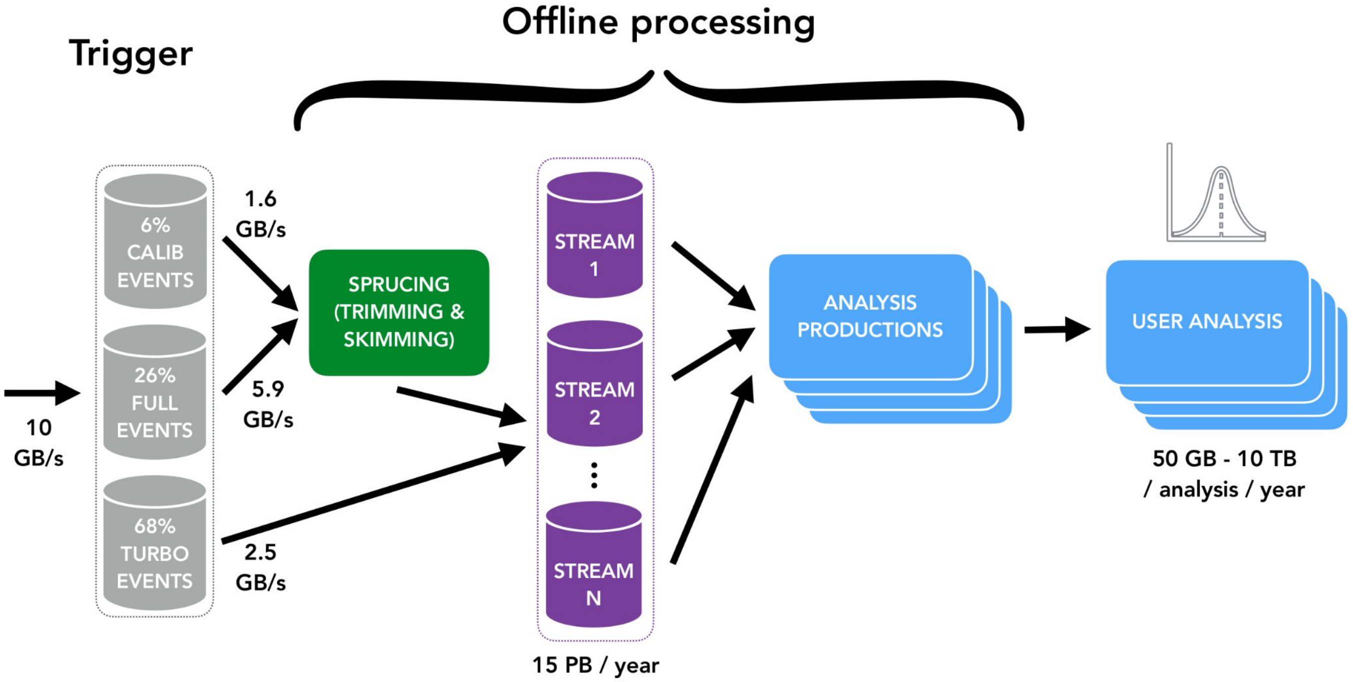
- There are many challenges to overcome related to open data releases at large experimental facilities
  - **Experiment side:** Very large data volumes! Need to either make a copy of the data (storage intensive) or provide external access (security risk)
  - **User side:** Large learning curve for using experiment specific software leads to high barrier of entry for external analysts
  - **Solution:** The Ntuple Wizard application offers a scalable solution to mitigate these problems
- Dataset discovery is motivated by choosing physics objects of interest
- Ntuples are configured with a web application in a user friendly way
- We recently submitted a paper to Computing and Software for Big Science
  - You can find it on the arxiv (https://arxiv.org/abs/2302.14235)

Summary (2): Ongoing Work

- We are working closely with CERN IT to get the Ntuple Wizard integrated with the CERN Open Data Portal!
- We are writing documentation to accompany use of the application
- Expecting a first release of the application towards the end of 2023! Stay tuned…
- User feedback will be welcome and appreciated! We are working on a system to handle this.

BACKUP

Run 3 Data Flow



LHCb-FIGURE-2020-016