Analysis Preservation & Open Data at LHCb

Ntuple Wizard and other efforts

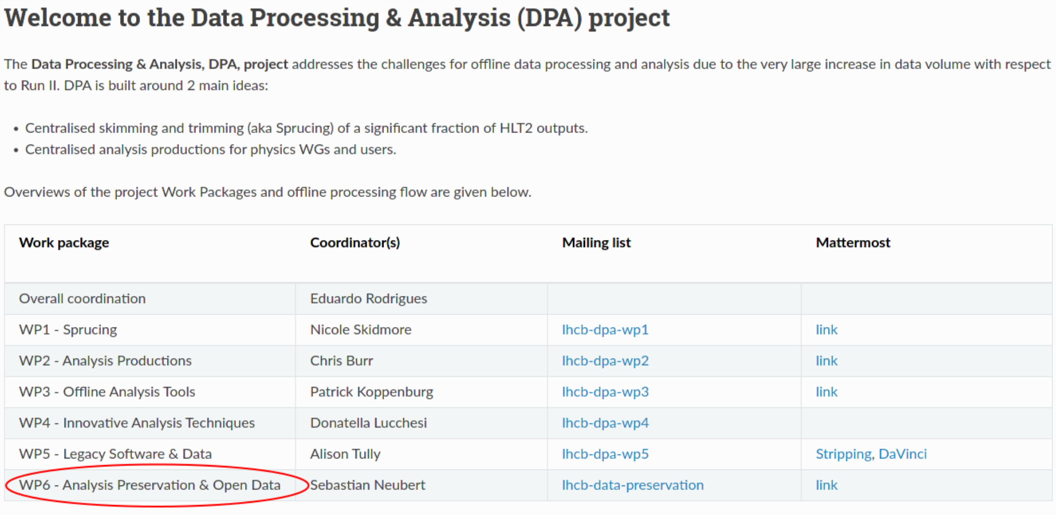
Dillon Fitzgerald[1], **Adam Morris**[2], on behalf of the LHCb AP&OD group

[1]University of Michigan, [2]University of Bonn

3rd DPHEP Collaboration Workshop, 2021-06-21

Organisation

Website: cern.ch/lhcb-dpa/wp6



Welcome to the Data Processing & Analysis (DPA) project

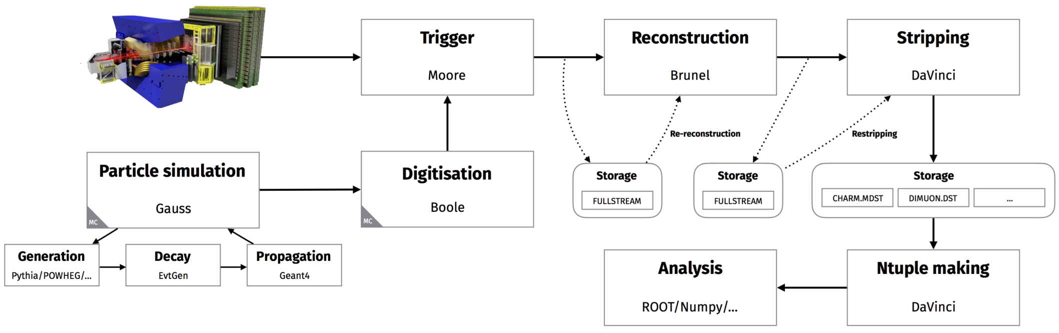
The **Data Processing & Analysis, DPA, project** addresses the challenges for offline data processing and analysis due to the very large increase in data volume with respect to Run II. DPA is built around 2 main ideas:

- Centralised skimming and trimming (aka Sprucing) of a significant fraction of HLT2 outputs.
- Centralised analysis productions for physics WGs and users.

Overviews of the project Work Packages and offline processing flow are given below.

Work package	Coordinator(s)	Mailing list	Mattermost
Overall coordination	Eduardo Rodrigues		
WP1 - Sprucing	Nicole Skidmore	lhcb-dpa-wp1	link
WP2 - Analysis Productions	Chris Burr	lhcb-dpa-wp2	link
WP3 - Offline Analysis Tools	Patrick Koppenburg	lhcb-dpa-wp3	link
WP4 - Innovative Analysis Techniques	Donatella Lucchesi	lhcb-dpa-wp4	
WP5 - Legacy Software & Data	Alison Tully	lhcb-dpa-wp5	Stripping, DaVinci
WP6 - Analysis Preservation & Open Data	Sebastian Neubert	lhcb-data-preservation	link

LHCb data processing in Run 1 (2011-2012)



- Data locations tracked with **Bookkeeping** system
- **Stripping:** build and select particle candidates, run centrally in periodic campaigns
  - **Stripping line:** specific candidate/selection requirement (c.f. trigger line)
- Analysis performed on ROOT **ntuples** derived from Stripping output
  - Created by running **user-submitted jobs** on the WLCG

Analysis Productions

User analysis jobs (old way)

Job ~ “process this file”

- ✗ Slow interface with WLCG (DIRAC)
- ✗ Manual rescheduling of failed jobs
- ✗ Options & output in user storage
  - ✗ Access depends on site availability
  - ✗ **Deleted when user leaves or needs to free up space**
- ✗ Testing manual and not enforced
  - ✗ many failed user jobs

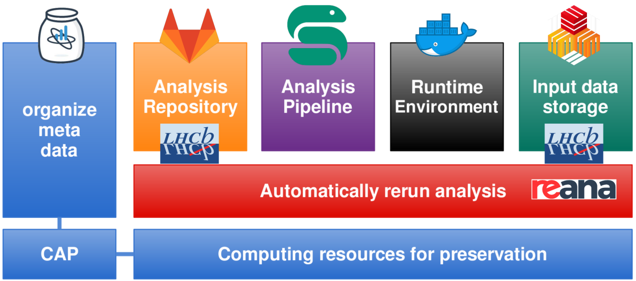
Analysis Productions (new way)

Production ~ “process this dataset”

- ✓ Web-based monitoring
- ✓ Same production system as Simulation, Reconstruction, Stripping
  - ✓ Jobs automatically (re)scheduled
- ✓ **Preservation of job options**
- ✓ Output stored centrally
- ✓ **Entry in Bookkeeping**
- ✓ Testing enforced before submission

Analysis preservation efforts

From the LHCb Analysis Preservation Roadmap (2017):



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

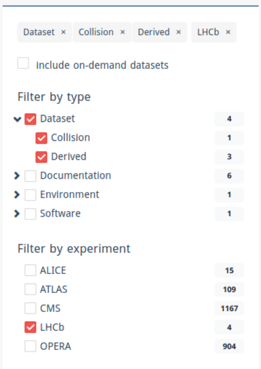
- Ongoing survey of current practices in the collaboration
  - How well is the current policy adopted?
  - How are workflows and environments being preserved now?
- Aim to make recommendations to LHCb management ~ end of year

Analysis preservation efforts

Technical developments & vested interests:

- Snakemake workflow template repo
- Support for Snakemake in REANA (CERN IT summer intern)
- Streamlined analysis lifecycle management
  - Merge patchwork of systems
  - Interface with CAP
- Tag-based access to Analysis Production output

LHCb Open Data: challenge #1



Dataset × Collision × Derived × LHCb ×

□ include on-demand datasets

Filter by type

☑ Dataset	4
☑ Collision	1
☑ Derived	3
□ Documentation	6
□ Environment	1
□ Software	1

Filter by experiment

□ ALICE	15
□ ATLAS	109
□ CMS	1167
☑ LHCb	4
□ OPERA	904

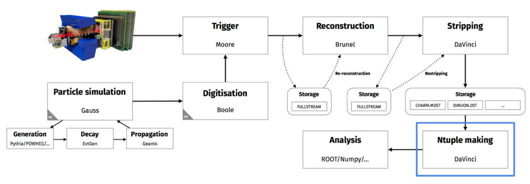
Where is the LHCb Open Data?

Working on a release of the Run 1 Stripping output…

Significantly larger filesizes than other LHC experiments!

Not scalable: quickly hit storage quota on Open Data Portal.

LHCb Open Data: challenge #2



Ntuple making currently requires knowledge of the LHCb software stack

- Lots of documentation and support required in order to be succesful
  - See LHCb Starterkit
- High barrier of entry for external analysts
- Complete overhaul of data processing (including ntupling) for Run 3
  - Knowledge required to access Run1+2 data will start fading “soon”

LHCb Open Data: a solution

Ntuple Wizard

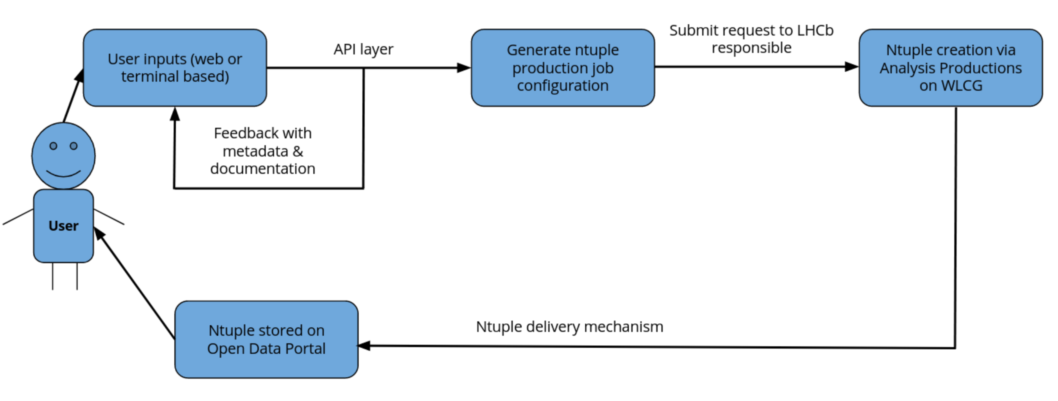
Mechanism for generation of ntuples without knowledge of the LHCb software stack

- Intuitive user inputs
- Produce necessary option files
- Create ntuples via Analysis Productions
- Return ntuples to user on the CERN Open Data Portal

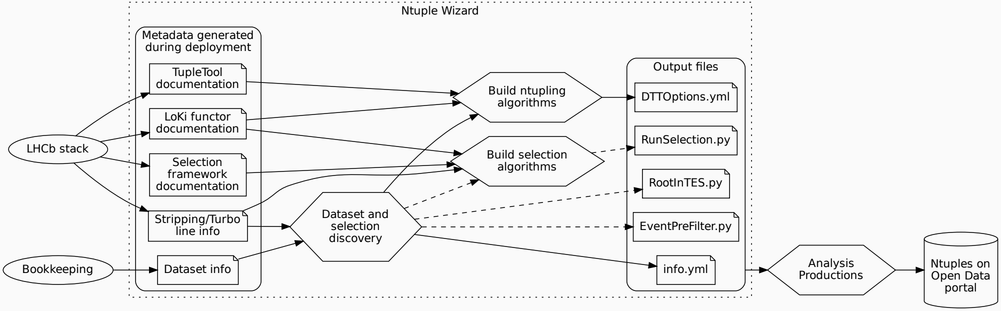
This solves several problems:

- ✓ Much smaller storage and bandwidth requirements on Open Data Portal
- ✓ Significantly flattened learning curve for accessing LHCb data
- ✓ No need for a computing cluster just to download and process the data

Overview


User inputs (web or terminal based)
API layer
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

In more detail


Ntuple Wizard
Metadata generated during deployment
TupleTool documentation
LoKi functor documentation
Selection framework documentation
Stripping/Turbo line info
Dataset info
LHCb stack
Bookkeeping
Dataset and selection discovery
Build ntupling algorithms
Build selection algorithms
Output files
DTTOptions.yml
RunSelection.py
RootInTES.py
EventPreFilter.py
info.yml
Analysis Productions
Ntuples on Open Data portal

Three core interactive components:

- Dataset and selection discovery
- Build selection algorithms
- Build ntupling algorithms

Development Strategy

1.) **Minimal working example** with basic functionality

- Focus on “build ntupling algorithms” component
- Limited to selecting already-built decay candidates

2.) Develop extensions in parallel:

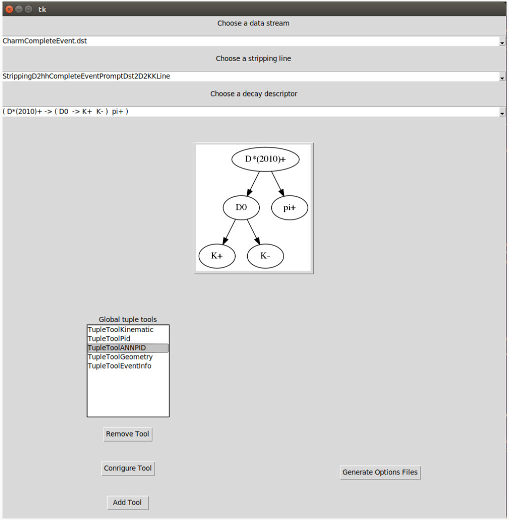
- **Web application** (summer student project)
  - Intuitive interface for algorithm configuration and input data selection
  - Display hints, examples and documentation
- **Dataset discovery** from more intuitive user inputs
- Configuration of **further algorithms**
  - selection sequences
  - mass hypothesis substitutions
  - custom jet reconstruction
  - etc…

Minimal Working Example

GUI for demonstration purposes.

To be replaced with web application

Video recording



Choose a data stream

CharmCompleteEvent.dst

Choose a stripping line

StrippingD2hhCompleteEventPromptDst2D2KKLine

Choose a decay descriptor

(D*(2010)+ -> (D0 -> K+ K-) pi+)

Global tuple tools

- TupleToolKinematic
- TupleToolPid
- TupleToolANNPID
- TupleToolGeometry
- TupleToolEventInfo

Considerations

Security & permissions

- ! DaVinci typically configured with python scripts
  - ✗ **Running arbitrary code from outside is a bad idea**
  - Wizard saves configuration as data structures, to be interpreted by our parsers
- ! Dataset discovery requires metadata from Bookkeeping and LHCb stack
  - Pull all metadata at “deployment time”
  - **Read static files at runtime**, no interaction with DIRAC
- ! Locations of “unreleased” datasets easily guessable
  - Check input data against allowed list
- ! LHCb policy reserves right to withhold part of dataset (e.g. ongoing analysis)
  - Require **fine-grained control** over:
    - building/accessing decay candidates
    - Stripping lines or equivalent selections
  - No elegant/agreed solution yet

Considerations

Limitations

- Aim to cover most common use-cases…
- … but cannot offer 100% of available functionality
  - Security/access implications
  - Development time

Summary

- Ntuple making from user jobs to central production system
- Lots of parallel developments towards more complete analysis preservation
- Open Data release is a challenge (size and learning curve)
  - Ntuple wizard offers a scalable solution

Thank you for listening