# Large-scale data processing and monitoring for run 3 with Analysis Productions at LHCb

Dylan Jaide White (they/them), on behalf of the LHCb collaboration

MANCHESTER 1824 The University of Manchester

## Overview and motivation

- The LHCb Upgrade I and the High-Luminosity LHC upgrade will increase both event rate and event size at LHCb during run 3 (starting in 2022)
  - Together, these will cause the data rate to increase by a factor of ~30× [1]
- Working with so much data comes with many challenges, including:
  - Data needs to be efficiently processed into analysis-ready formats
  - New code for run 3 needs to be monitored regularly, so that any problems with collected data can be found and fixed as soon as possible
  - Everything needs to be stored in an organised manner
  - The system to do all of this must be easy for analysts to learn and use
- **Analysis Productions** was designed to solve all of these problems

## The run 3 LHCb detector and data flow

- Single-arm forward spectrometer, specialised for studying beauty and charm hadrons
- Almost every part of the detector has been upgraded from run 2, to handle the higher event rate
- Data Processing and Analysis (DPA) project is developing upgrades to offline data handling

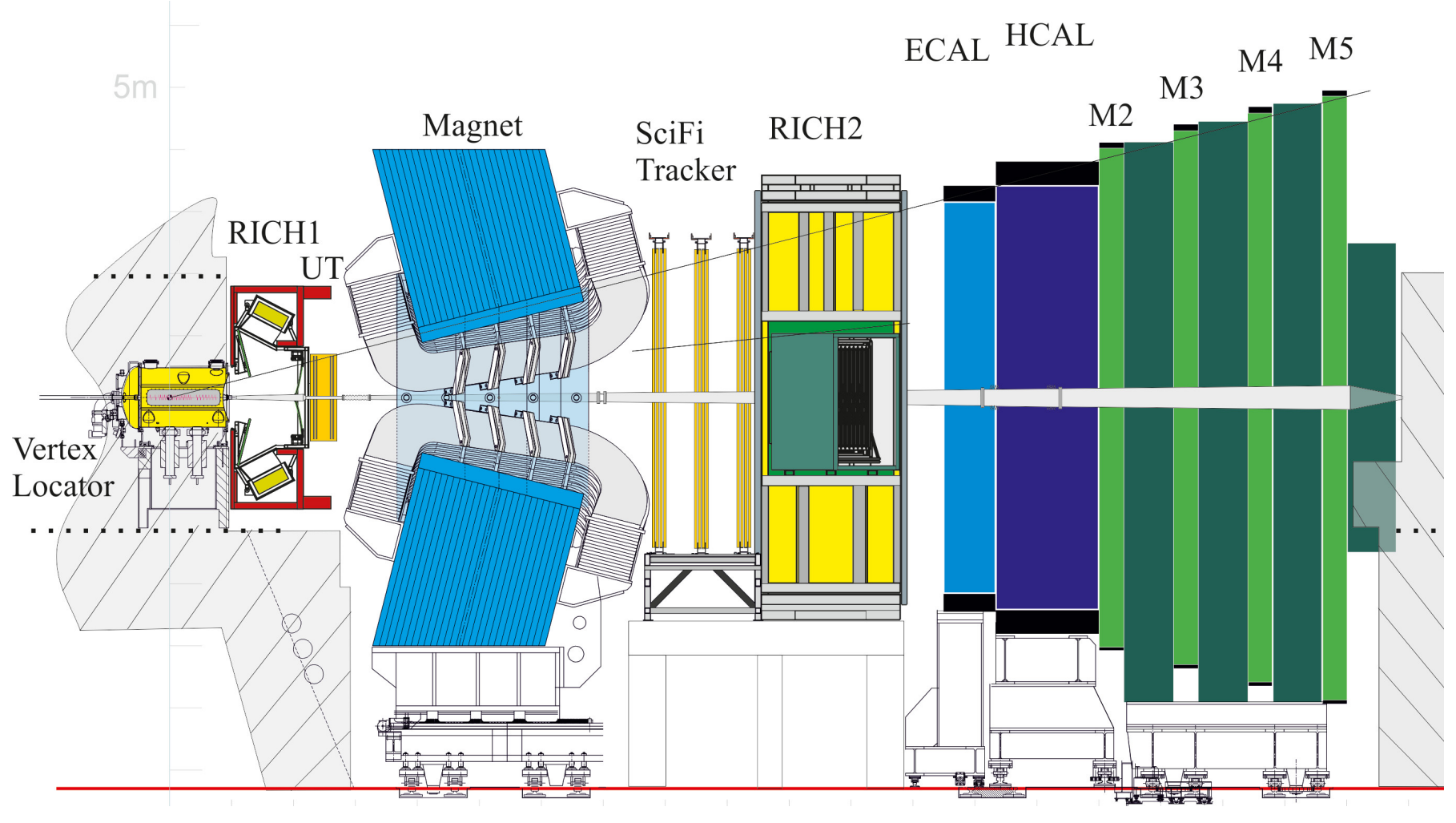
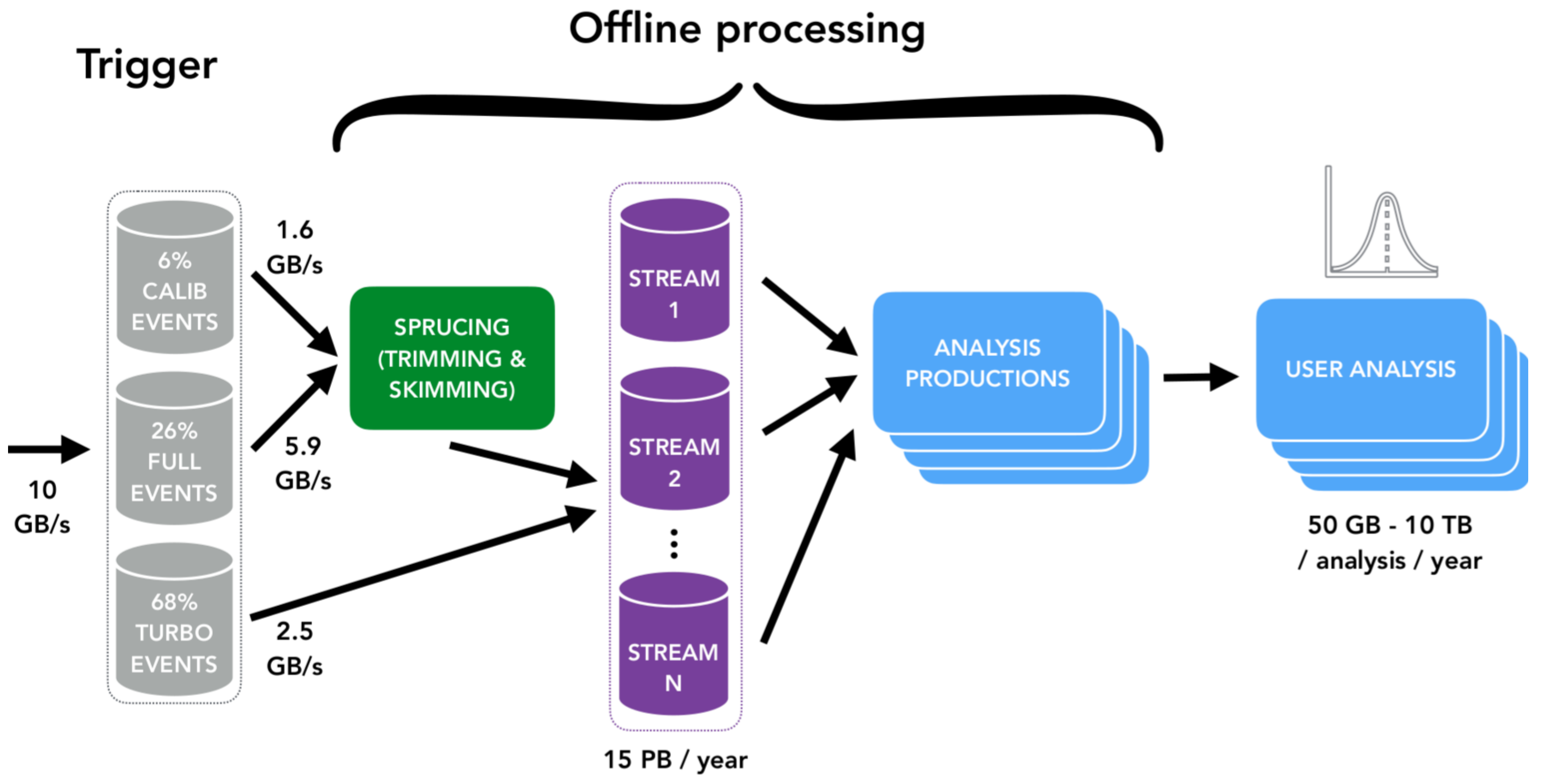


**Figure:** The LHCb detector for run 3

- Final step of processing is putting data into analysis-ready ntuples
  - Expected data rates of $\mathcal{O}(10)$ petabytes per year



Trigger

Offline processing

10 GB/s

6% CALIB EVENTS

26% FULL EVENTS

68% TURBO EVENTS

1.6 GB/s

5.9 GB/s

2.5 GB/s

SPRUCING (TRIMMING & SKIMMING)

STREAM 1

STREAM 2

⋮

STREAM N

15 PB / year

ANALYSIS PRODUCTIONS

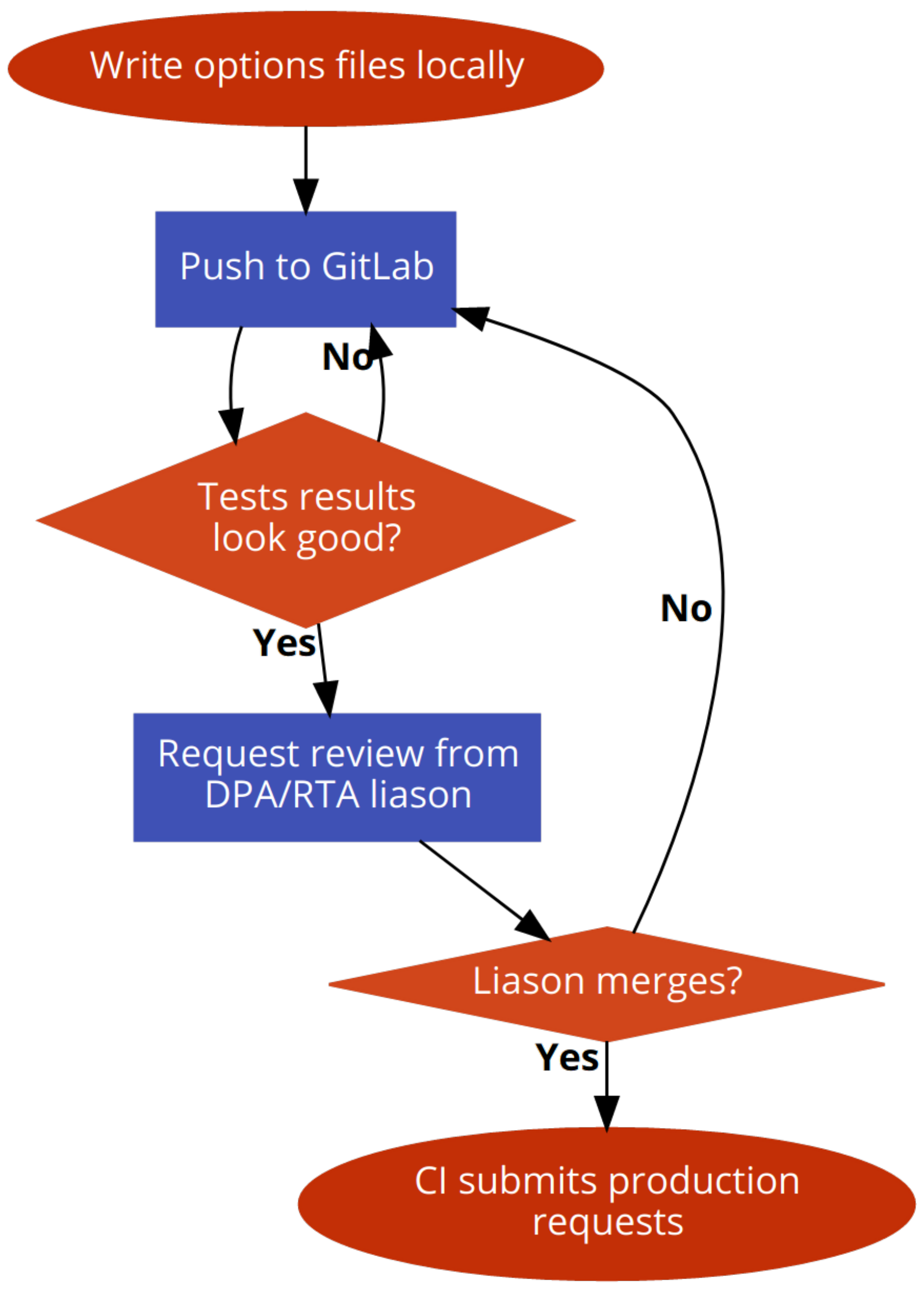
USER ANALYSIS

50 GB - 10 TB / analysis / year

**Figure:** Dataflow for LHCb during run 3 [2]

## Analysis Productions

- For most of runs 1-2 (2011-2018), everyone had to make their own ntuples
  - Most users needed to become familiar with an intricate system to submit ntuple-making jobs to batch computing resources manually
  - Any mistakes often required full re-runs, wasting computing time
  - Code and ntuples were stored in user areas, making them difficult to preserve



Write options files locally

Push to GitLab

No

Tests results look good?

Yes

Request review from DPA/RTA liason

No

Liason merges?

Yes

CI submits production requests

**Figure:** Analysis Productions workflow

- The Analysis Productions system improves on this [3]
  - Automated job submission and handling of failed jobs
  - Jobs are validated before submission by automatic CI tests and liaisons
  - Ntuples are stored centrally and logged in bookkeeping, code is persisted on GitLab
- To create a new production, users need to supply only two things:
  - Job options file
  - YAML configuration file for Analysis Productions
- This has now been used by most working groups, and works well
- Combined with the new analysis software framework, creation of ntuples is easier and more robust.

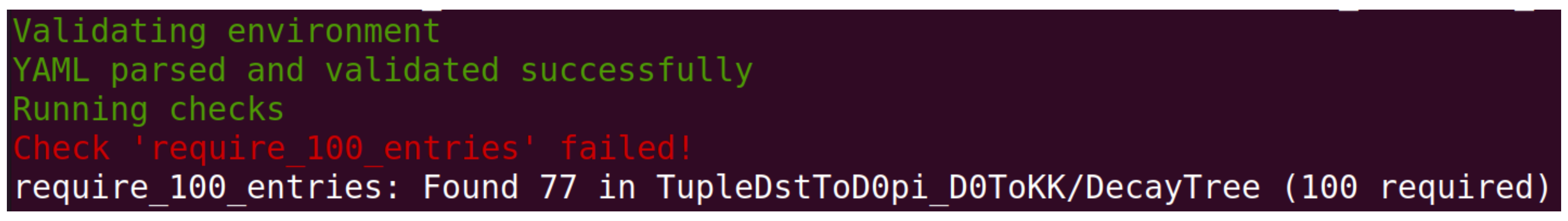
## Improving validation with checks

- Automated validation performed with CI tests is able to easily catch mistakes that cause errors or crashes
- But up to now, it has had no way to detect issues with data quality
  - This has so far been done manually, by analysts and liaisons
  - But this takes lots of effort, and mistakes are easy to make
- The new **checks** feature aims to solve this
- Users can define checks in a production's YAML configuration file

```yaml
checks:
  require_100_entries:
    type: num_entries
    count: 100
```

**Figure:** YAML syntax for configuring a simple check

- Checks are run automatically when testing productions, including in CI tests
- If a check fails, the production won't be approved until this is fixed



```
Validating environment
YAML parsed and validated successfully
Running checks
Check 'require_100_entries' failed!
require_100_entries: Found 77 in TupleDstToD0pi_D0ToKK/DecayTree (100 required)
```

**Figure:** Output from a failed check

- There are many types of checks, including some which create histograms and perform simple background subtraction on spectra of interest
  - All results are displayed in the Analysis Productions Web App

## Using checks for offline monitoring

- With the checks feature, Analysis Productions can automatically process stored data and create custom histograms
- This means it can be used to create automated offline monitoring plots
- Aim: to allow analysts to regularly check their data early in run 3
  - Help to catch any mistakes so they can be fixed sooner rather than later
- Users can easily configure monitoring to suit their own analyses
  - Uses the same files that they will later need for a full production
  - Supports blinding of data in all checks which create histograms

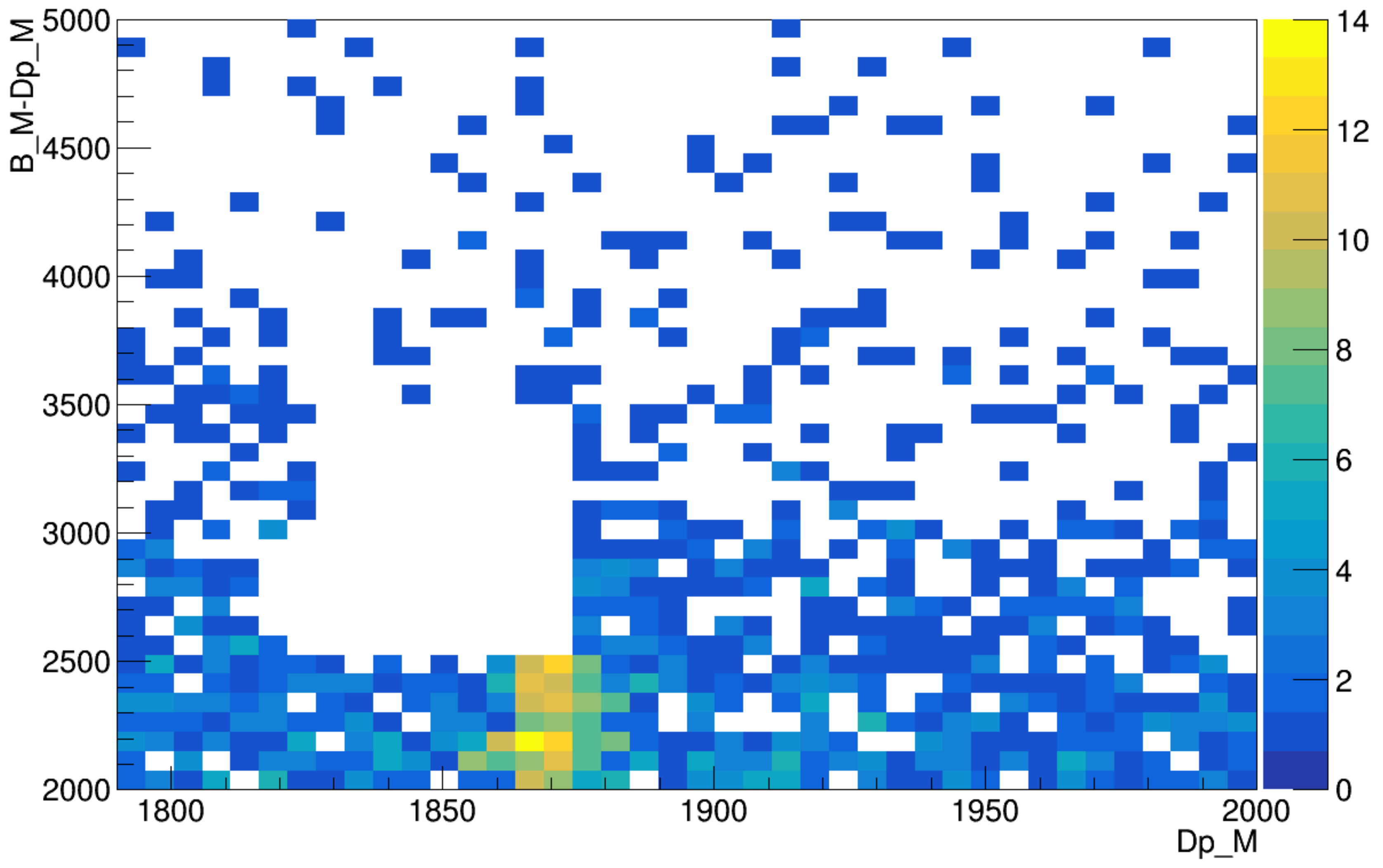


**Figure:** 2D histogram with blinding, produced by a check

## Summary

- The Analysis Productions system has been created to improve the experience of creating ntuples with robust built-in validation
- Checks have been added to help automate the validation of data quality
- This is being developed further into an offline monitoring tool for run 3

## References

[1] CERN/LHCC 2018-14, 2018.

[2] LHCb-FIGURE-2020-016, 2020.

[3] N. Skidmore et al. Run-3 offline data processing and analysis at LHCb, PoS (EPS-HEP2021) 792, 2022.

View this poster on your device: