Belle II Data production

2024 Belle Analysis Workshop

Renu
on behalf of the Data Production group

19th October, 2024 - 23th October, 2024

Belle II Data Production

- **Primary Goal:**

  - Smooth, timely production of data and MC samples for physics analysis and other studies
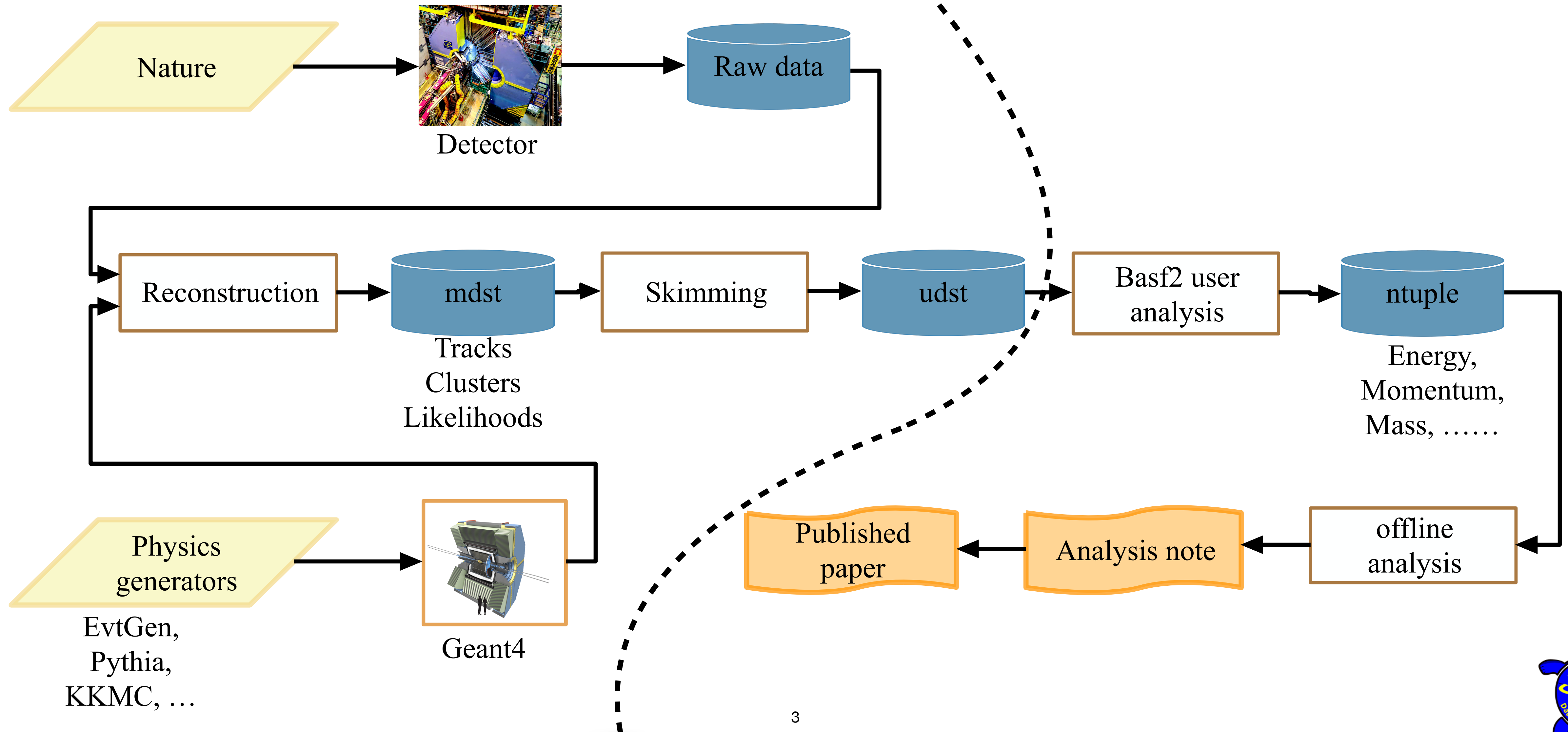
- **Tasks:**

  - Calibration and alignment (https://confluence.desy.de/display/BI/Data+Production+Calibration)

  - RAW data (re)processing (https://confluence.desy.de/display/BI/Phase+3+data)

  - MC production (https://confluence.desy.de/display/BI/Data+Production+MC12)

  - Analysis skimming (https://confluence.desy.de/display/BI/Skimming+Homepage)

The Big Picture

Managed by collaboration

Managed by analyst

Nature → Detector → Raw data

Reconstruction → mdst → Skimming → udst → Basf2 user analysis → ntuple → offline analysis

mdst: Tracks, Clusters, Likelihoods

ntuple: Energy, Momentum, Mass, ……

Physics generators: EvtGen, Pythia, KKMC, … → Geant4

Published paper ← Analysis note ← offline analysis

3

The Big Picture

Data Confluence page

[Data Production WebHome](#)

Data Confluence page

[Data Production WebHome](#)

Contact details
for DP leadership

Who's who and contacts

Coordinators: @Stefano Lacaprara @gdujany (deputy) - Former: @tamponi @jbennett

Skim manager: @scavino @vberta @shdelamo

Calibration manager: @Renu Garg , @tamponi (deputy)

Validation manager: @charris @pecker

Data processing manager: @pgironel @watanuki (deputy),

MC processing manager:, @Giovanni Gaudino @Gaurav Sharma (deputy)

MC signal: @hanx @liwenzhe

Operation manager: @gdujany

DP leadership responsibilities are listed here. (To be updated)

Meetings and Mailing list
Mailing list: dataprod@belle2.org

Meetings: meetings page.

Minutes: https://hackmd.io/0oKpJkveRhWdJ5uQ6y3__g

(2023) https://hackmd.io/vGeJNbMqSV6F01cY3bVZ0w

(2022) https://hackmd.io/dbskL9vDQjeQ1PXuu-Zzog

Shift Manuals

- Standard DP shift manual
- DC Expert shift manual

Data production liaisons
(responsibilities of the data production liaisons can be found here)

Group	Liaison
Semileptonic and Missing Energy Decays	@Tommy Martinov @charris
Radiative & Electroweak Penguin	@iprudiev
Time Dependent CP Violation	@Noah Brenny
Hadronic B decays	@Xiaodong Shi
Quarkonium	@liyang71 @zhulin
Charm	@Jaeyoung Kim
Tau	@naveen10 @Kenji Inami
Dark-sector and low multiplicity	@Gaurav Sharma
Performance	@lekaiyao
Upgrade	@benni

Physics working group
liaisons (collect MC
production requests,
perform validation
tasks, etc.)

Technical pages of the DP sub-groups

Here you can find information relevant for DP members. Nothing needed by analysis should end up here

Other resources

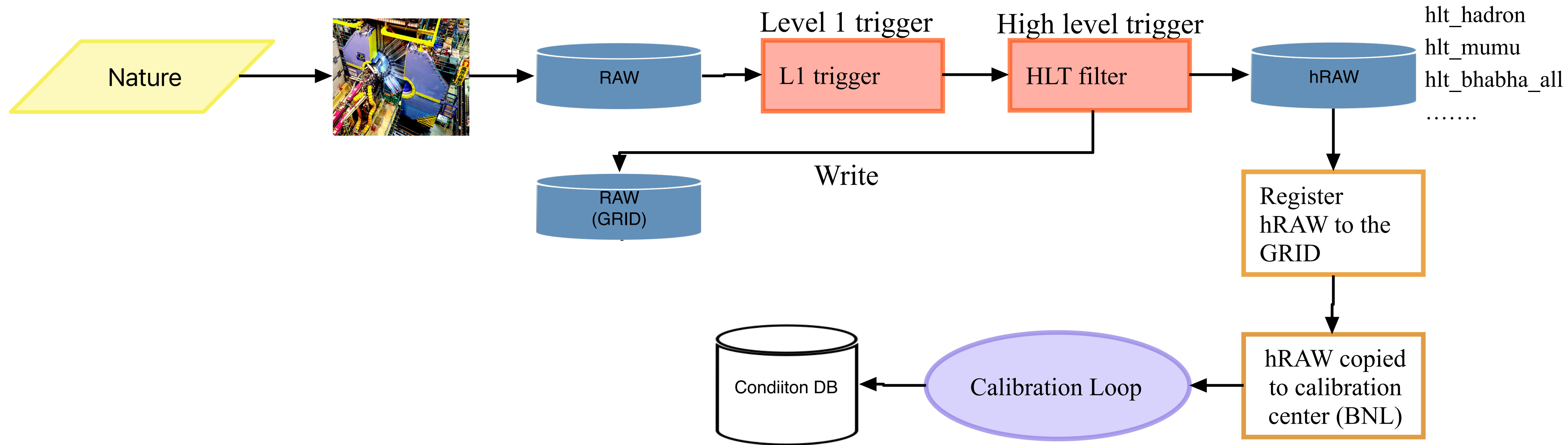Check this page for the description of the data formats (RAW, mDST, cDST, uDST) and their usage

Data format

In general, Belle II output is stored in ROOT files containing various subsets of dataobjects, dbobjects, nTuples, etc

RAW:	un-processed, un-calibrated output of the detector
hRAW:	same as RAW, but only for events passing a given HLT filter or skim
DST	Data Summary Table - All available dataobjects are included (not generally used as it contains everything)
cDST:	calibration Data Summary Table - cDST contain RAW data and additional data-objects useful for calibration
mDST:	mini Data Summary Table - Controlled version of a DST. - only a subset of available processed data-objects are included - Flagged skim approach! - Use for most analysis (see below)
uDST:	user Data Summary Table - mDST objects plus analysis objects (e.g. particleLists) - produced from skims - reduce time needed to run analysis jobs - Samples created only for FEI based analysis!!

Data Flow

Nature → RAW → Level 1 trigger [L1 trigger] → High level trigger [HLT filter] → hRAW

hlt_hadron
hlt_mumu
hlt_bhabha_all
…….

Write → RAW (GRID)

Register hRAW to the GRID

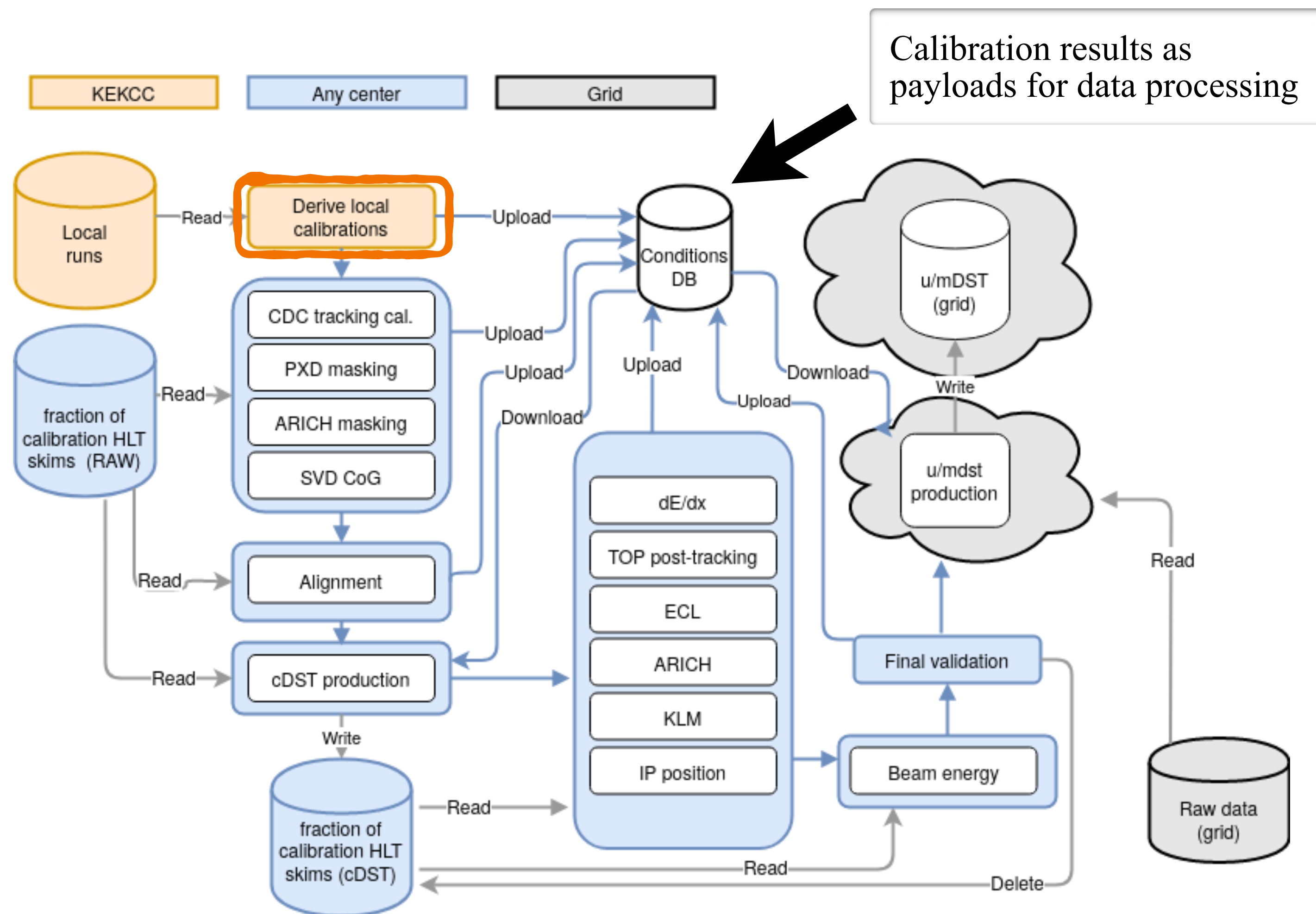hRAW copied to calibration center (BNL)

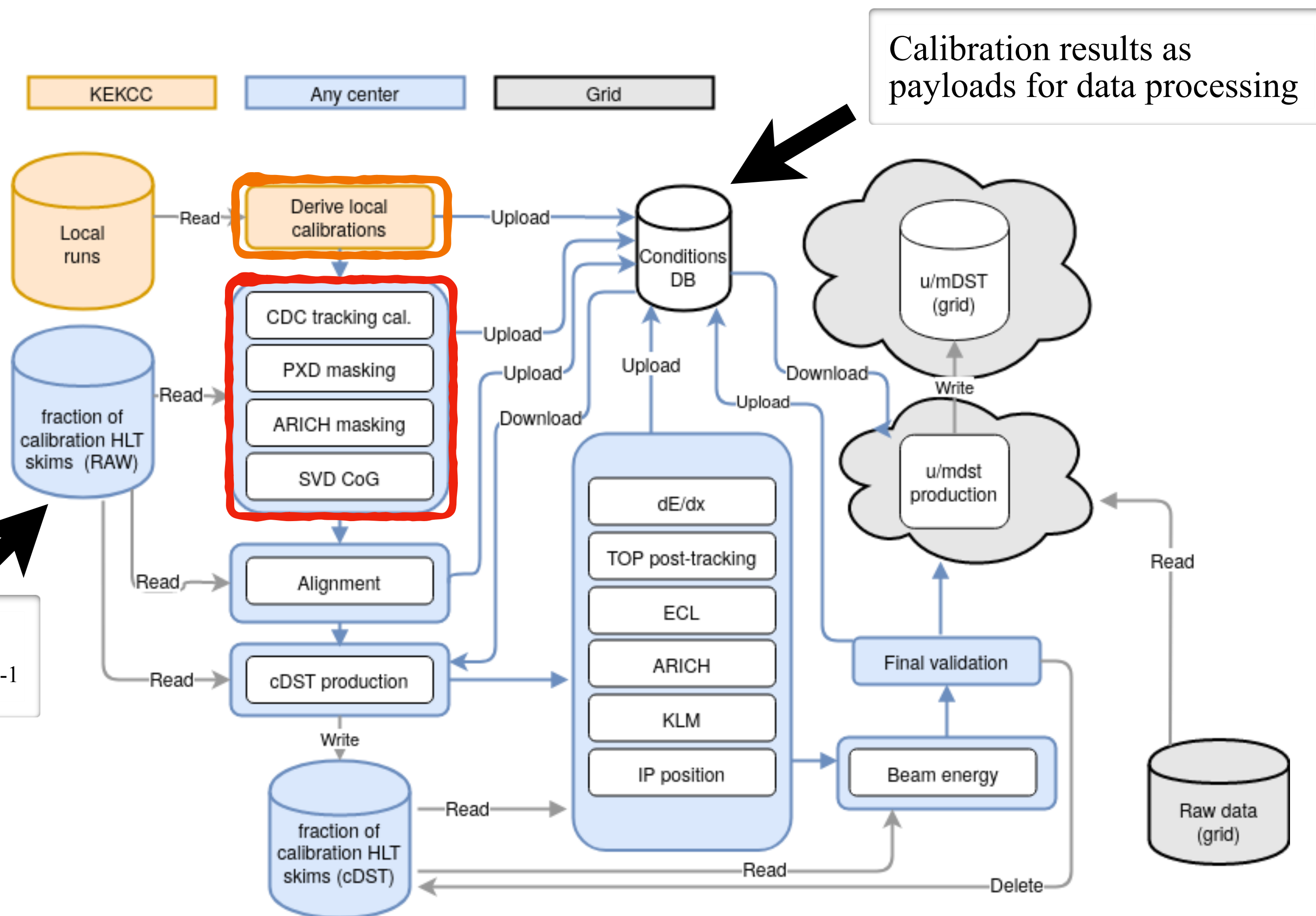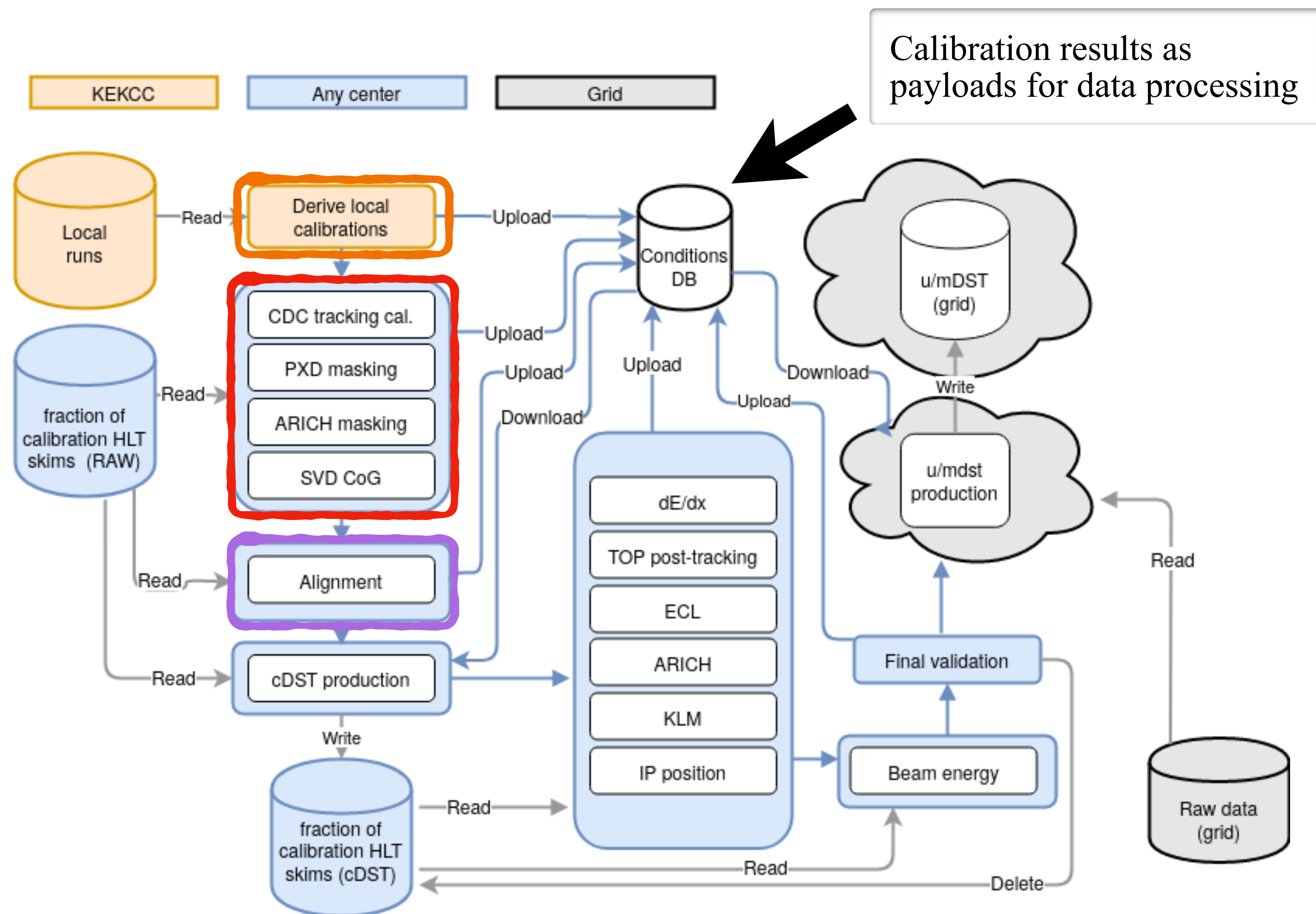Calibration Loop

Condiiton DB

Calibration Loop

The **goal** of the calibration is to provide data usable for physics analysis

Full calibration loop divided into 5 steps:

- **Local calibration:** derived from local runs or DQM (e.g TOP laser calibration, SVD noise calibration)

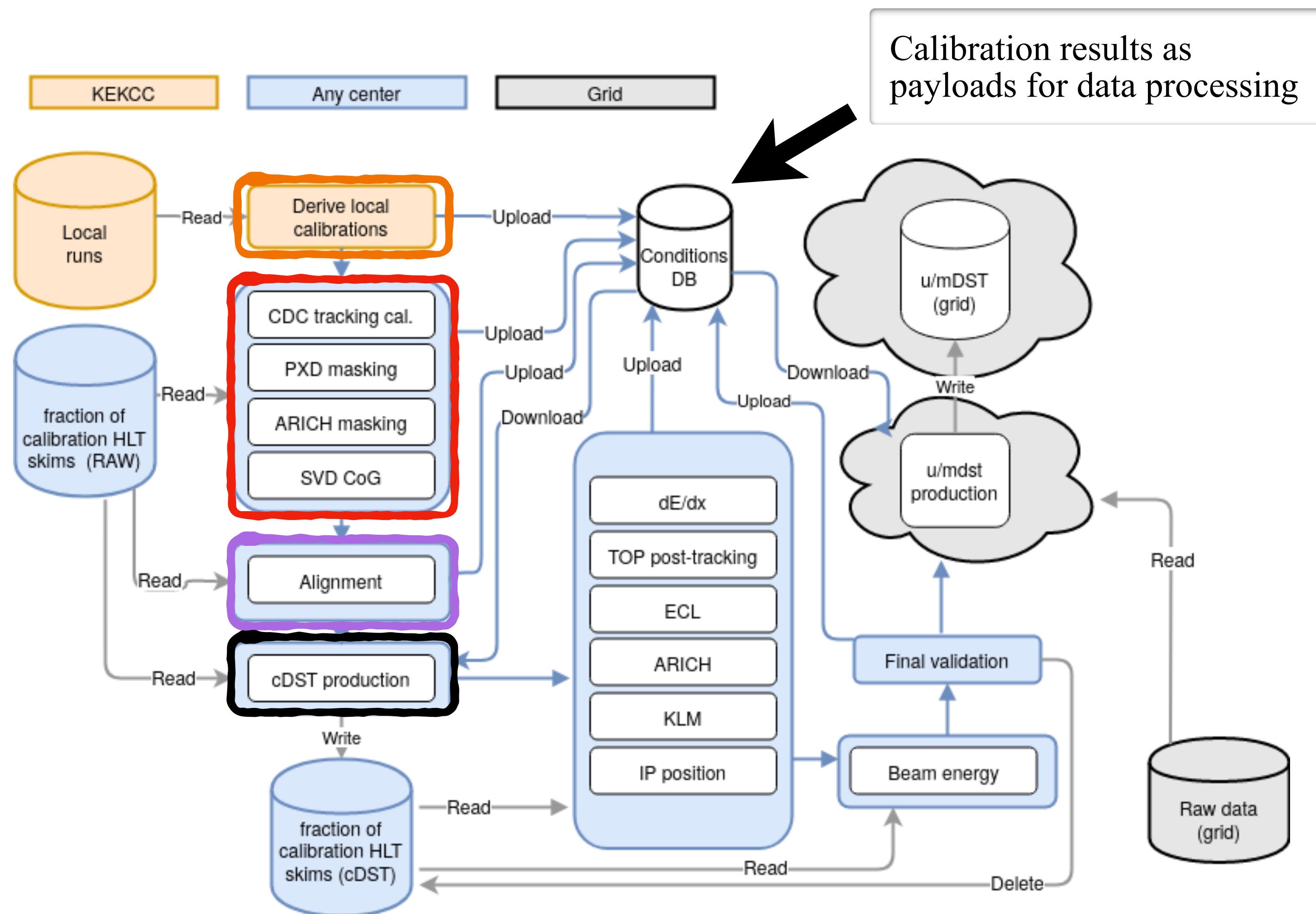Calibration results as payloads for data processing

Calibration Loop

The **goal** of the calibration is to provide data usable for physics analysis

Full calibration loop divided into 5 steps:

- **Local calibration:** derived from local runs or DQM (e.g TOP laser calibration, SVD noise calibration)

- **Raw data based calibration:** must run on raw collision data (e.g channel masking, CDC tracking calibration)

Calibration results as payloads for data processing

Calibration skims (hRAW)
Adaptive pre-scaling to 12fb^{-1}
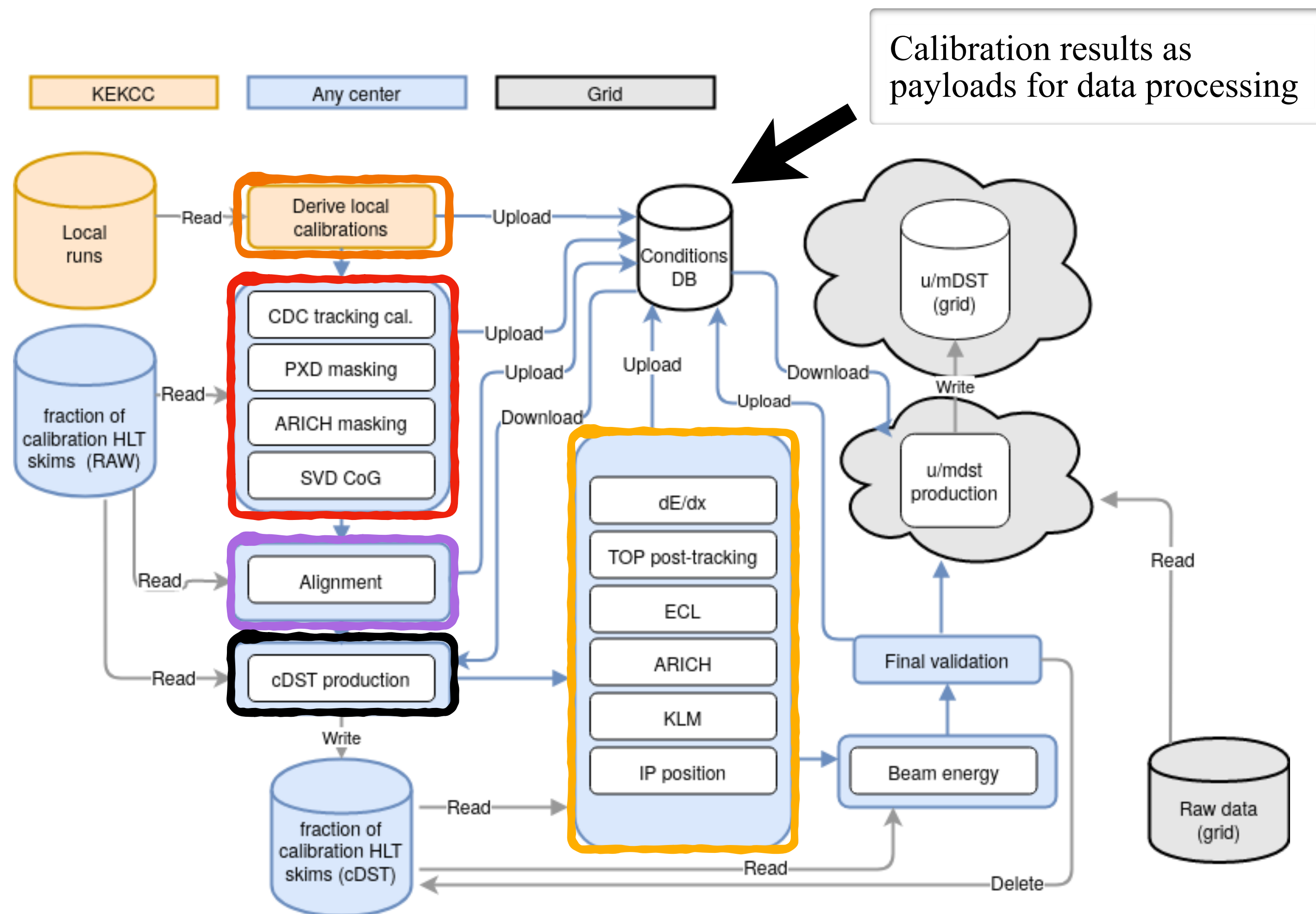
Calibration page

Calibration Loop

The **goal** of the calibration is to provide data usable for physics analysis

Full calibration loop divided into 5 steps:

- **Local calibration:** derived from local runs or DQM (e.g TOP laser calibration, SVD noise calibration)

- **Raw data based calibration:** must run on raw collision data (e.g channel masking, CDC tracking calibration)

- **Alignment:** requires raw collision data and both local and the other raw data based calibrations

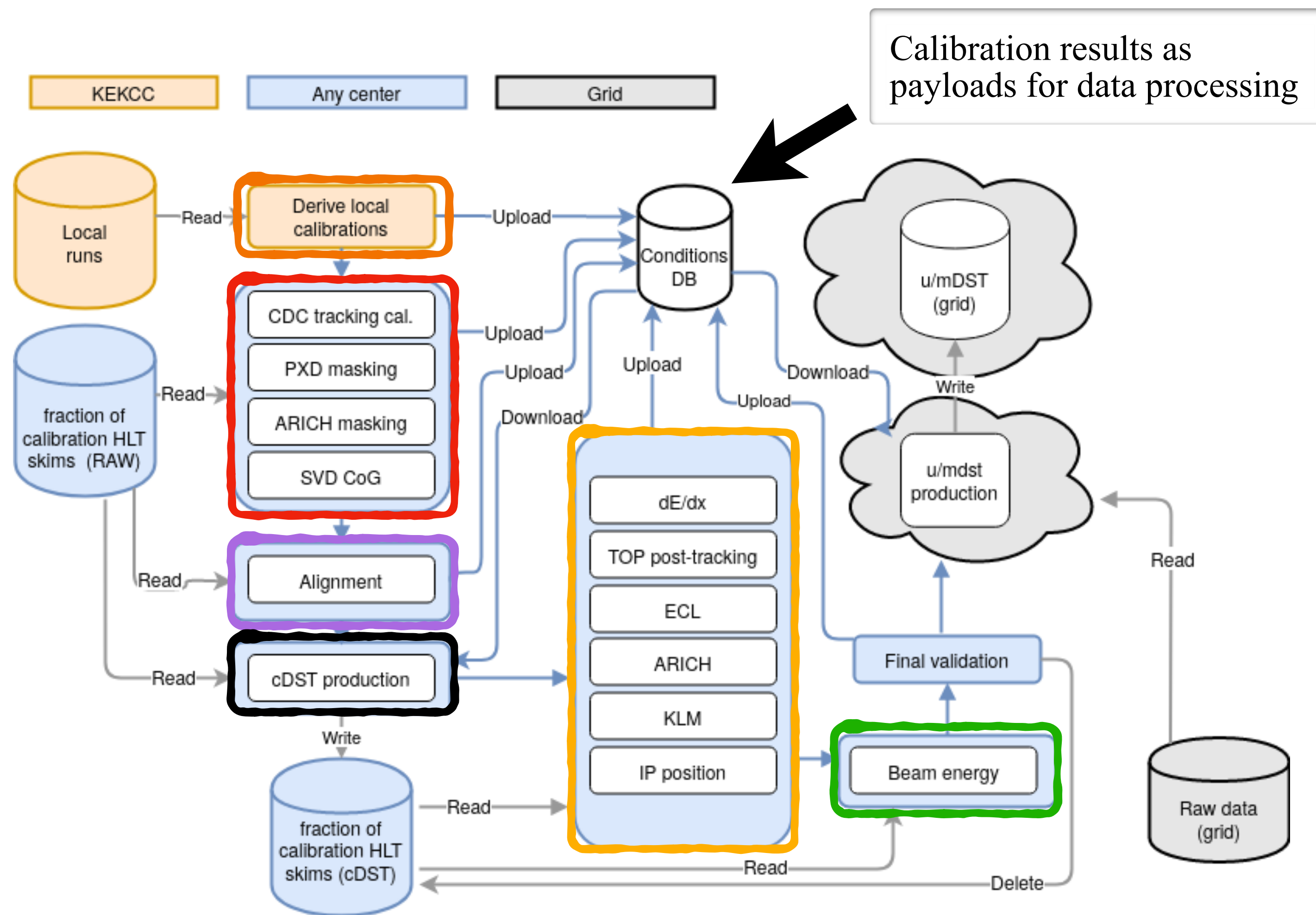
Calibration results as payloads for data processing

Calibration Loop

The **goal** of the calibration is to provide data usable for physics analysis

Full calibration loop divided into 5 steps:

- **Local calibration:** derived from local runs or DQM (e.g TOP laser calibration, SVD noise calibration)

- **Raw data based calibration:** must run on raw collision data (e.g channel masking, CDC tracking calibration)

- **Alignment:** requires raw collision data and both local and the other raw data based calibrations

- **cDST production**

Calibration results as payloads for data processing

Calibration Loop

The **goal** of the calibration is to provide data usable for physics analysis

Full calibration loop divided into 5 steps:

- **Local calibration:** derived from local runs or DQM (e.g TOP laser calibration, SVD noise calibration)

- **Raw data based calibration:** must run on raw collision data (e.g channel masking, CDC tracking calibration)

- **Alignment:** requires raw collision data and both local and the other raw data based calibrations

- **cDST production**

- **Post-tracking calibration:** require good tracks, run on centrally produced cDST files (e.g CDC dE/dx)

Calibration results as payloads for data processing

Calibration Loop

The **goal** of the calibration is to provide data usable for physics analysis

Full calibration loop divided into 5 steps:

- **Local calibration:** derived from local runs or DQM (e.g TOP laser calibration, SVD noise calibration)

- **Raw data based calibration:** must run on raw collision data (e.g channel masking, CDC tracking calibration)

- **Alignment:** requires raw collision data and both local and the other raw data based calibrations

- **cDST production**

- **Post-tracking calibration:** require good tracks, run on centrally produced cDST files (e.g CDC dE/dx)

- **Analysis based calibration:** rely on high quality data (e.g ECMS calibration)

Calibration Example

Run gain is calculated as fit mean of dE/dx distribution from each run (bhabha tracks)
- **Run gain changes because instabilities of CDC**
  - ‣ Gas composition (incl. water vapor)
  - ‣ Temperature
  - ‣ Pressure

Run Gain (e^+e^-)

dE/dx resolution for bhabha's ~ 7%
Variation versus run gain ~ 7%

March 28, 2024

April 30, 2024

2601

33 days

3508

15

Calibration Confluence page

[Calibration page](#)

Global Tag

IoV — **Interval of Validity**
An experiment and run interval for which a payload is valid.

Payload — An atom of condition data (e.g CDCDedxRunGain)
Identified by name and revision number
ROOT format

Global tag — Collection of payloads and their IoVs
Has a unique name and a description

IoVs

- consists of four values:
  first_exp, first_run, final_exp, final_run
  e.g: 30, 24, 30, 140

Special case:

- final_exp>=0 && final_run<0:
  valid for all runs in final_exp
  e.g: 30, 0, 30, -1

- final_exp <0 && final_run<0:
  valid forever
  e.g: 30, 0, -1, -1

Correct global tags are automatically selected during processing

Global Tag

Condition database: https://cdbweb.sdcc.bnl.gov

Questions? Write to *T1 CDB team*

| Home | **GlobalTag** | Payload | Types of Payload | Global Tag Comparison |

GlobalTag: 3194 items found   Click on items for more details

ID: [ ]   Name (can be partial): [ ]   Status: [All ▼]   Type: [All ▼]   Modified by: [ ]   items per page: [25 ▼]   [Submit]

| Name ▲	ID ▲	Default? ▲	Description ▲	Status ▲	Type ▲	Modified ▲	Mod. by ▲	Total Payloads ▲	Dist
mcrd_prompt_rel08	3308	⊖	Globaltag that contains the simulation payloads for MC16rd to superseed the ones in data_prompt_rel08 and online	TESTING	DEV	05/03/2024 10:12 a.m.	gdujany	0	0
mcrd_proc16	3307	⊖	Globaltag that contains the simulation payloads for MC16rd to superseed the ones in data_proc16 and online	TESTING	DEV	05/03/2024 8:28 a.m.	gdujany	1622	43
user_gaudino_mcrd_proc16	3306	⊖	staging MCrd GT for proc16 production	TESTING	DEV	05/03/2024 8:02 a.m.	gaudino	1622	43
temp_gdujany_beam_parameters_proc16_chunk1	3305	⊖	Beam parameters for proc16 chunk1	OPEN	DEV	04/29/2024 1:55 p.m.	gdujany	1566	1
neutrals_2024	3304	⊖	GT with photon energy bias correction variation and photon efficiency corrections obtained with preLS1-data and MC15rd	TESTING	DEV	04/26/2024 10:14 a.m.	eganiev	21	3
user_cwessel_ROISimulationParameters_update	3303	⊖	Special ROISimulationParameters for exp12 runs 1188 and 1189, c.f. gitlab issue 10473.	OPEN	DEV	04/25/2024 7:55 a.m.	cwessel	1	1
AIRFLOW_operation_staging_proc16_chunk2	3302	⊖	staging Gt for chunk2	OPEN	DEV	04/24/2024 2:37 p.m.	tamponi	0	0
user_takaham_pxd_exp1004_5deadL2modules	3301	⊖	user GT for PXD with PXDDeadPixelPar including 5 L2 modules which have been OFF since the beginning of Run2 physics run	TESTING	DEV	04/21/2024 12:27 p.m.	takaham	1	1
user_lacaprar_test3_software_trigger_cut	3300	⊖	A test GT to upload only software_trigger_cut from online	OPEN	DEV	04/19/2024 3:24 p.m.	lacaprar	108	108
user_lacaprar_test2_software_trigger_cut	3299	⊖	A test GT to upload only software_trigger_cut from online	OPEN	DEV	04/19/2024 3:23 p.m.	lacaprar	14	14
user_lacaprar_test_software_trigger_cut	3298	⊖	A test GT to upload only software_trigger_cut from online	TESTING	DEV	04/19/2024 3:11 p.m.	lacaprar	108	108
		⊖	Globaltag for storing payloads						

(Re)Processing scheme

For any given data, two types of calibration and processing happens:

Prompt processing:

- Calibration @ BNL
- Minimum luminosity: 12 fb^{-1}
- Uses hRAW as input
- All calibrations included
- Terminology: **bucketXX**
- ~one bucket every 2 weeks of data taking
- Ideally already final calibration

Official reprocessing:

- Recalibration @ KEKCC
- To update older data with the latest reconstruction software
- Uses cDST as input
- Only calibration with expected improvement
- Terminology: **procXX**
- ~once/year until 2025, every other year starting from 2025

Single bucket $\geq$ 12 fb^{-1}
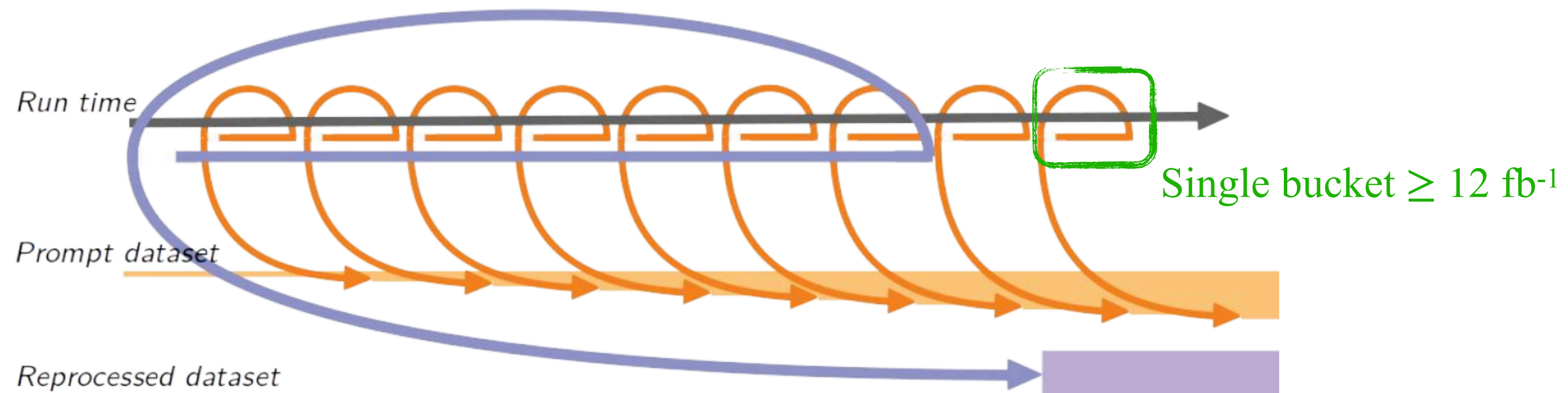
(Re)Processing scheme

For any given data, calibration and processing happens twice:

Prompt processing:

- Calibration @ BNL

- Minimum luminosity: 12 fb^{-1}

- Uses hR

- All calib

- Termin

- ~one bu

- Ideally a

Official reprocessing:

- Recalibration @ KEKCC

- To update older data with the latest

mprovement

er year starting

Current data campaign:

Prompt processing:
- **Bucket37-40**: experiment 30, 31, 32, 33
- Integrated luminosity: 103.63 fb^{-1}

Official processing: Done
- **Proc16**: experiments ~~7-8, 10,~~ 12, 14, 16-18, 20-22, 24-26
- Integrated luminosity: 442.8 fb^{-1}

Single bucket $\geq$ 12 fb^{-1}

Prompt dataset

Reprocessed dataset
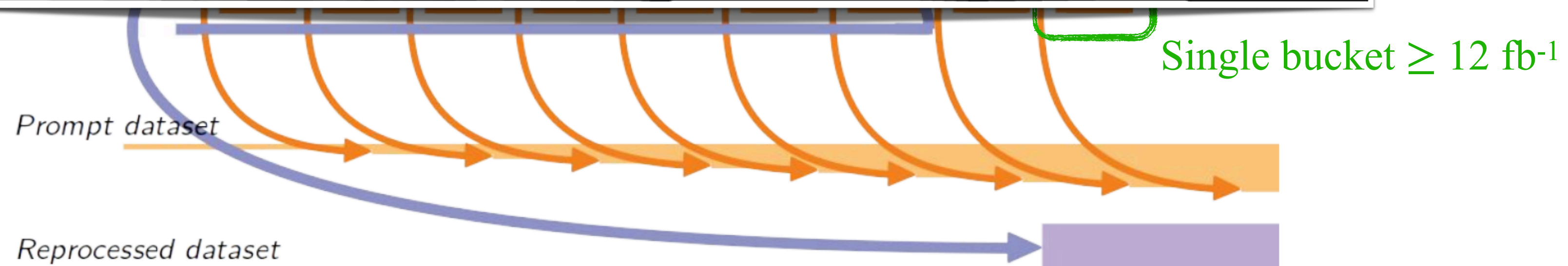
Data Flow
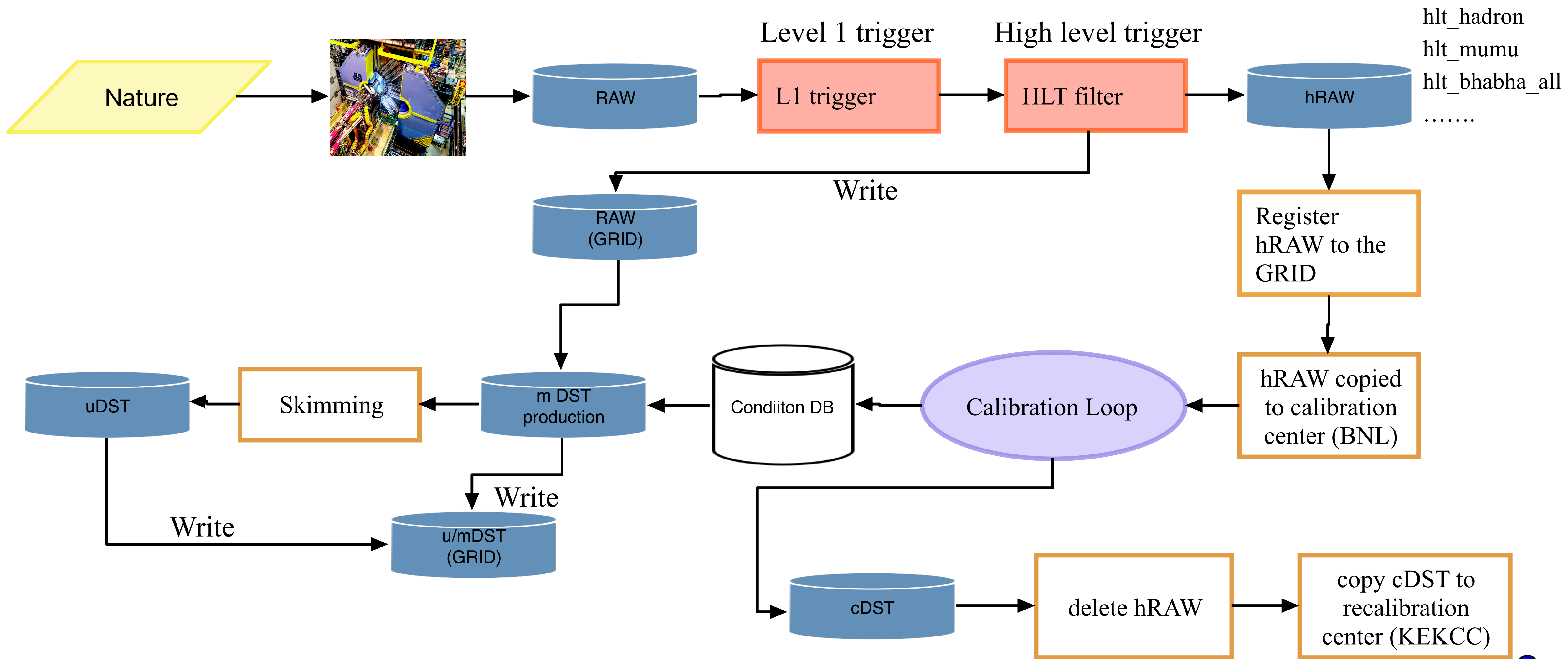
Analysis Skim

Skims are meant to provide analysis-oriented MC and data in reduced sizes

- Produced as uDST = mDST information + analysis-level information
  - Particle objects: which links particle hypotheses with tracks, neutral clusters, and particle identification information
  - Vertex fit results
  - Information from full B and D reconstruction, Continuum suppression and other complex algorithms
- **That means more information in smaller files!**

- This allows for preprocessing that reduces the CPU requirements for the analysis
- Your grid jobs on skimmed samples will finish much more quickly!
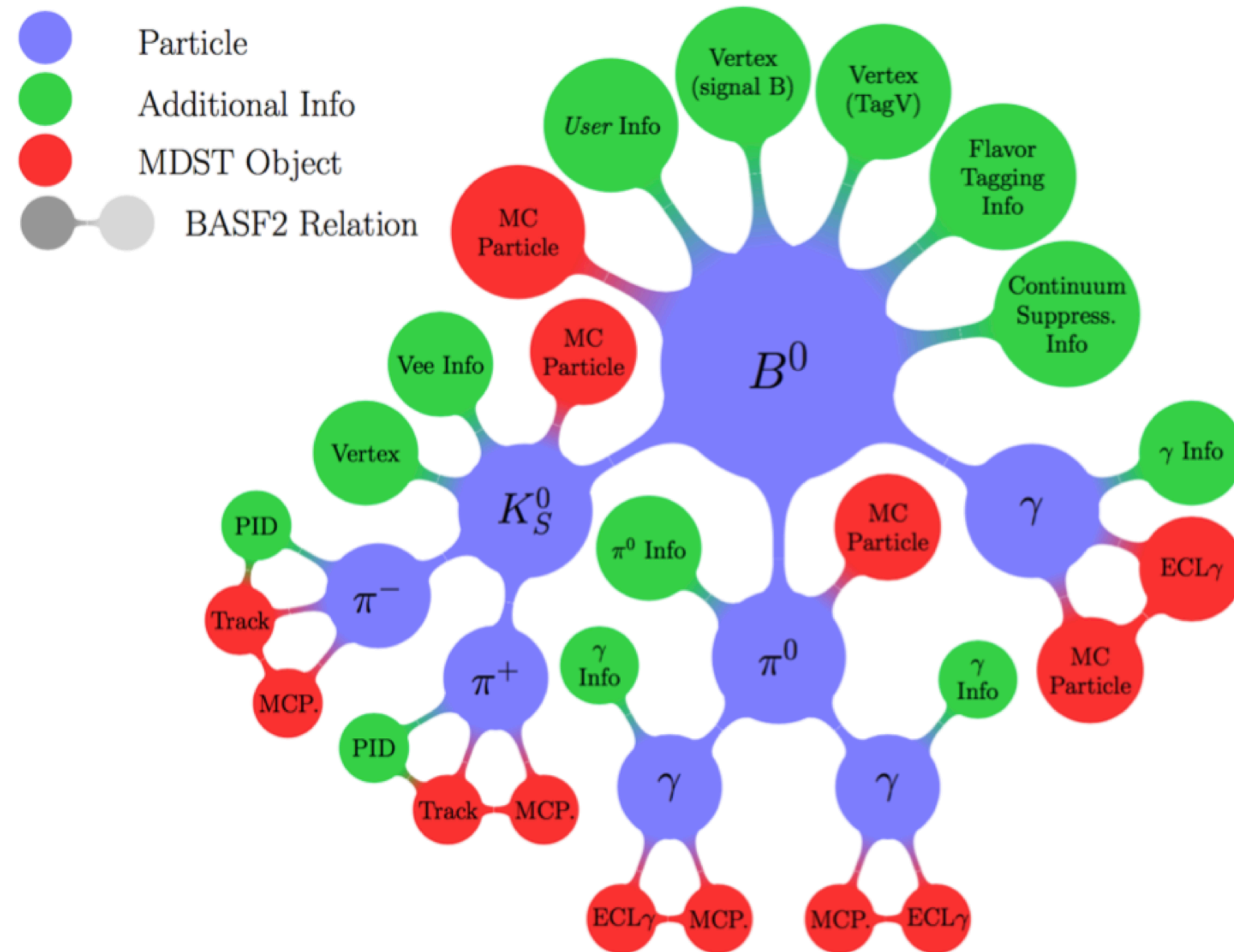
Use uDST for FEI based analysis!!

Flagged skim approach

Starting with proc16 and MC16 campaign a Flagged skim approach will be used

For more details:

- **Merged** different skims into groups
- Add a **flag** to each event according to the skim, to identify the skim within the group
- Produce **mDST** output

- **Why?**
  - Reduce the amount of work
  - Exploit the large overlap between the skims
  - Make submissions faster and lighter

- **How to use?**
  - Need to add the flag selection when you make the ntuples

```
ma.inputMdstList(filelist=infile, path=path)
ma.applyEventCuts(
        'eventExtraInfo(passes_mySkim)==1',
        path=path)
```

Use flagged mDST for your analysis, if not FEI based!!
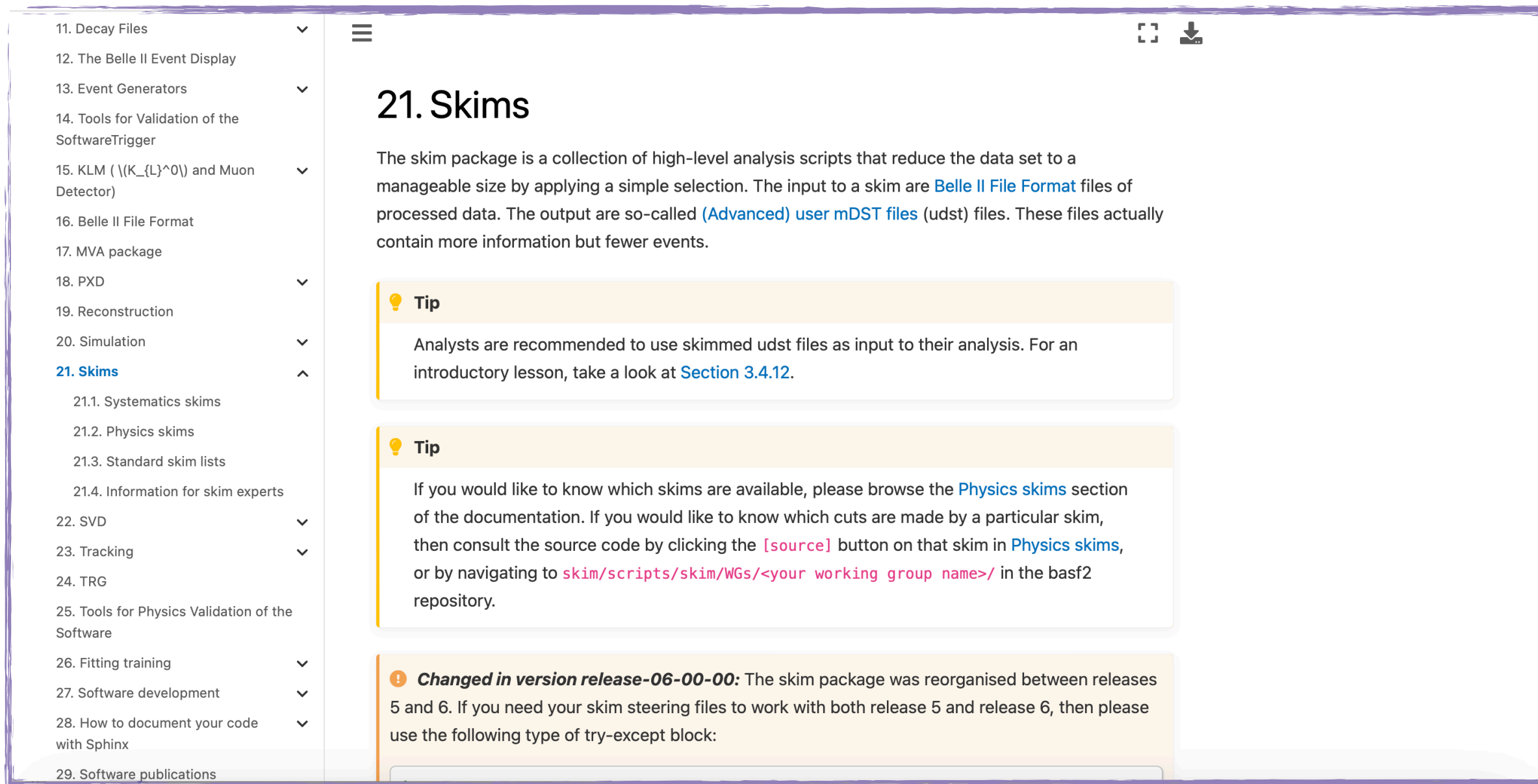
Understanding of Skimming

Two primary sources for information:

Documentation

- Good for general information and tutorials

Source code on gitlab

- Full information and always up to date

It is important for analysts to understand the skim they are using to make sure that it is aligned with their analysis goals.

Finding your Skim

Three ways to find your skims:

1. **Dataset searcher on DIRAC (web version)**

   - Provide a list of LFNs to use with gbasf2 analysis

   - https://dirac.cc.kek.jp:8443/DIRAC/

   - DIRAC apps →Dataset searcher

2. **Dataset searcher via command line (gbasf2 environment)**

   - source /cvmfs/belle.kek.jp/grid/gbasf2/pro/bashrc

   - gb2_ds_search dataset --data_type data -- campaign proc13 --exp_low 17 —exp_high 17 -- general_skim hadron

https://gbasf2.belle2.org/

Finding your Skim

3. Collections

- The easiest way to process data or MC as an analyst!

  - Contains the full list of LPNs for a given dataset

  - Ensures you use the correct files and don't miss any

  - Available for skims (currently only by request...)

https://gbasf2.belle2.org/collectionSearcher.html
Collection Summary, Data collection, MC collection

Skim Confluence page

Skim page

MC production

Two types of MC production:

- **Signal MC:**
  - Specific to your own analysis
  - **Dec files:** specify your own dec file according to the dec file naming rules
  - Contact the **Data production liaison** in your working group to get started!

- **Generic MC:**
  - Produced automatically at every MC production campaigns
  - They are just the typical processes which we expect to see at Belle II, such as:

  $e^+e^- \to \Upsilon(4S) \to B^+B^-$ (charged), B^0B^0 (mixed)
  $e^+e^- \to u\bar{u}, d\bar{d}, c\bar{c}, s\bar{s}$
  $e^+e^- \to \gamma\gamma, e^+e^-, \mu^+\mu^-, \tau^+\tau^-$ (taupair)
  $e^+e^- \to \ell\ell XX$ ($ee\pi\pi$, $eepp$, etc.), hhISR ($\pi\pi$ISR, KKISR, etc.)

  - Generated based on central decay file: DECAY_BELLE2.DEC

Data production liaisons
(responsibilities of the data production liaisons can be found here)

Group	Liaison
Semileptonic and Missing Energy Decays	@Tommy Martinov @charris
Radiative & Electroweak Penguin	@iprudiev
Time Dependent CP Violation	@Noah Brenny
Hadronic B decays	@Xiaodong Shi
Quarkonium	@liyang71 @zhulin
Charm	@Jaeyoung Kim
Tau	@naveen10 @Kenji Inami
Dark-sector and low multiplicity	@Gaurav Sharma
Performance	@lekaiyao
Upgrade	@benni

MC production campaign

Two types of MC production campaign:

Run-independent (RI)

- Easier to produce

- Use simulated background and static detector conditions

- Produced in predetermined luminosity

- Less accurate detector performance and beam backgrounds.
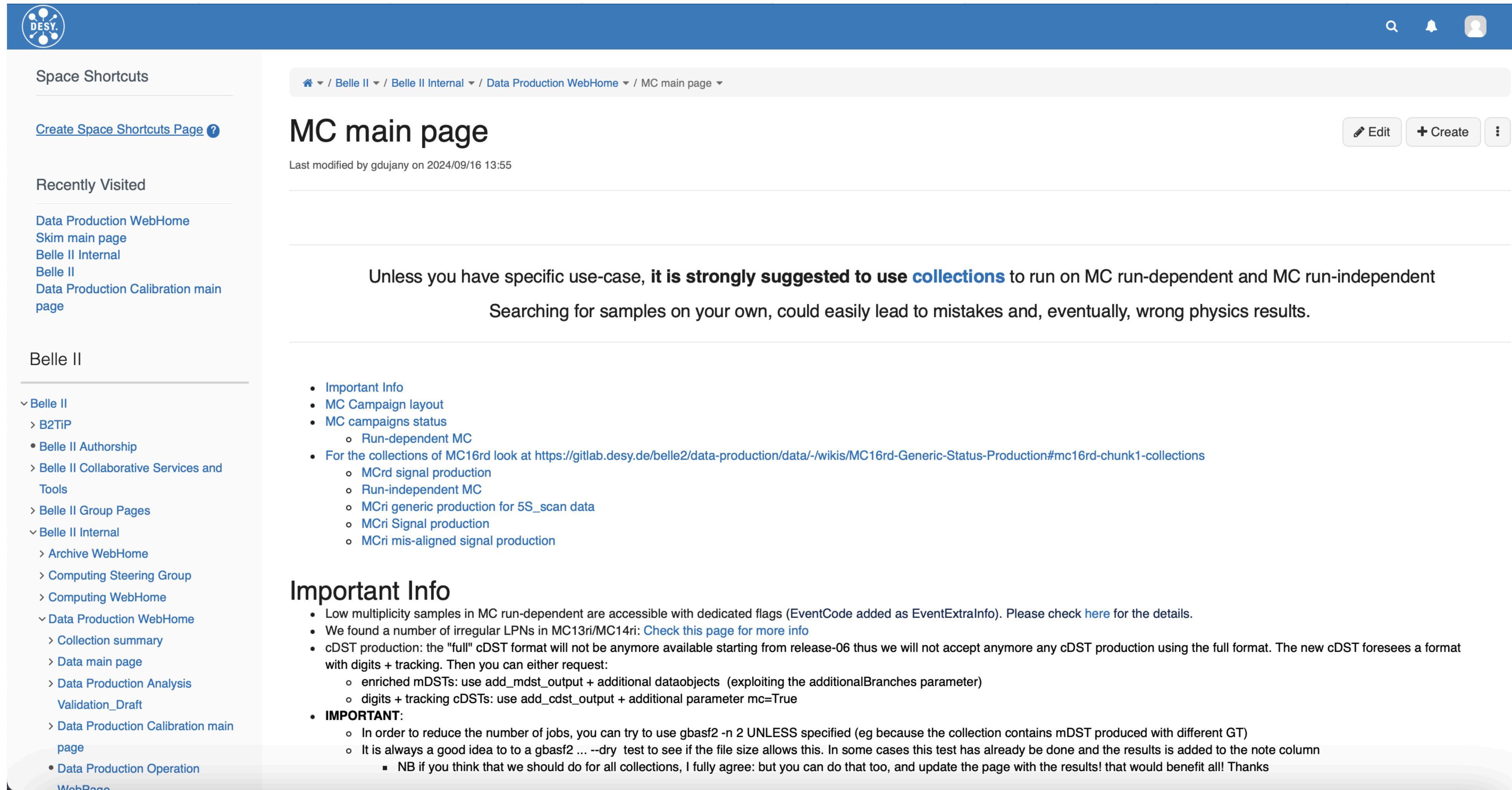
- Terminology: **MC16ri_X**

Run-dependent (RD)

- Difficult to produce

- Use random trigger events from data and real conditions

- Produced in streams (1 stream = luminosity of corresponding data)

- More accurate detector performance and beam backgrounds

- Terminology: **MC16rd_X**

MC Confluence page

MC page

Space Shortcuts

Create Space Shortcuts Page ❓

Recently Visited

Data Production WebHome
Skim main page
Belle II Internal
Belle II
Data Production Calibration main page

Belle II

- Belle II
  - B2TiP
  - Belle II Authorship
  - Belle II Collaborative Services and Tools
  - Belle II Group Pages
  - Belle II Internal
    - Archive WebHome
    - Computing Steering Group
    - Computing WebHome
    - Data Production WebHome
      - Collection summary
      - Data main page
      - Data Production Analysis Validation_Draft
      - Data Production Calibration main page
    - Data Production Operation WebPage

🏠 ▾ / Belle II ▾ / Belle II Internal ▾ / Data Production WebHome ▾ / MC main page ▾

MC main page

Last modified by gdujany on 2024/09/16 13:55

✏ Edit   + Create   ⋮

Unless you have specific use-case, **it is strongly suggested to use collections** to run on MC run-dependent and MC run-independent

Searching for samples on your own, could easily lead to mistakes and, eventually, wrong physics results.

- Important Info
- MC Campaign layout
- MC campaigns status
  - Run-dependent MC
- For the collections of MC16rd look at https://gitlab.desy.de/belle2/data-production/data/-/wikis/MC16rd-Generic-Status-Production#mc16rd-chunk1-collections
  - MCrd signal production
  - Run-independent MC
  - MCri generic production for 5S_scan data
  - MCri Signal production
  - MCri mis-aligned signal production

Important Info

- Low multiplicity samples in MC run-dependent are accessible with dedicated flags (EventCode added as EventExtraInfo). Please check here for the details.
- We found a number of irregular LPNs in MC13ri/MC14ri: Check this page for more info
- cDST production: the "full" cDST format will not be anymore available starting from release-06 thus we will not accept anymore any cDST production using the full format. The new cDST foresees a format with digits + tracking. Then you can either request:
  - enriched mDSTs: use add_mdst_output + additional dataobjects (exploiting the additionalBranches parameter)
  - digits + tracking cDSTs: use add_cdst_output + additional parameter mc=True
- **IMPORTANT**:
  - In order to reduce the number of jobs, you can try to use gbasf2 -n 2 UNLESS specified (eg because the collection contains mDST produced with different GT)
  - It is always a good idea to to a gbasf2 ... --dry test to see if the file size allows this. In some cases this test has already be done and the results is added to the note column
    - NB if you think that we should do for all collections, I fully agree: but you can do that too, and update the page with the results! that would benefit all! Thanks

More great resources

- https://www.belle2.org/

- **chat:** https://chat.belle2.org

- **questions:** https://questions.belle2.org

- **sympa (email lists):** https://lists.belle2.org/sympa/home

- **Mailing list:** dataprod@belle2.org, dataprod-skim@belle2.org, software-calibration@belle2.org

- **Basf2 documentation (Sphinx):** https://software.belle2.org/

- **gbasf2:** https://confluence.desy.de/display/BI/Computing+GBasf2
  - https://confluence.desy.de/display/BI/Instructions+for+gbasf2+analysis

- **Data production:** https://confluence.desy.de/display/BI/Data+production+WebHome

- **Gitlab:** https://gitlab.desy.de/belle2/

- **Experiment Numbering:** https://confluence.desy.de/display/BI/Experiment+numbering

- **Conditions Database:** https://cdbweb.sdcc.bnl.gov/ (globaltag information)

- **DIRAC (for dataset searcher):** https://dirac.cc.kek.jp:8443/DIRAC/