



Belle II Data production



2024 Belle II Summer Workshop

Renu on behalf of the Data Production group Supported by US DOE funding 17th June, 2024 - 21st June, 2024

4

Belle II Data Production

• Primary Goal:

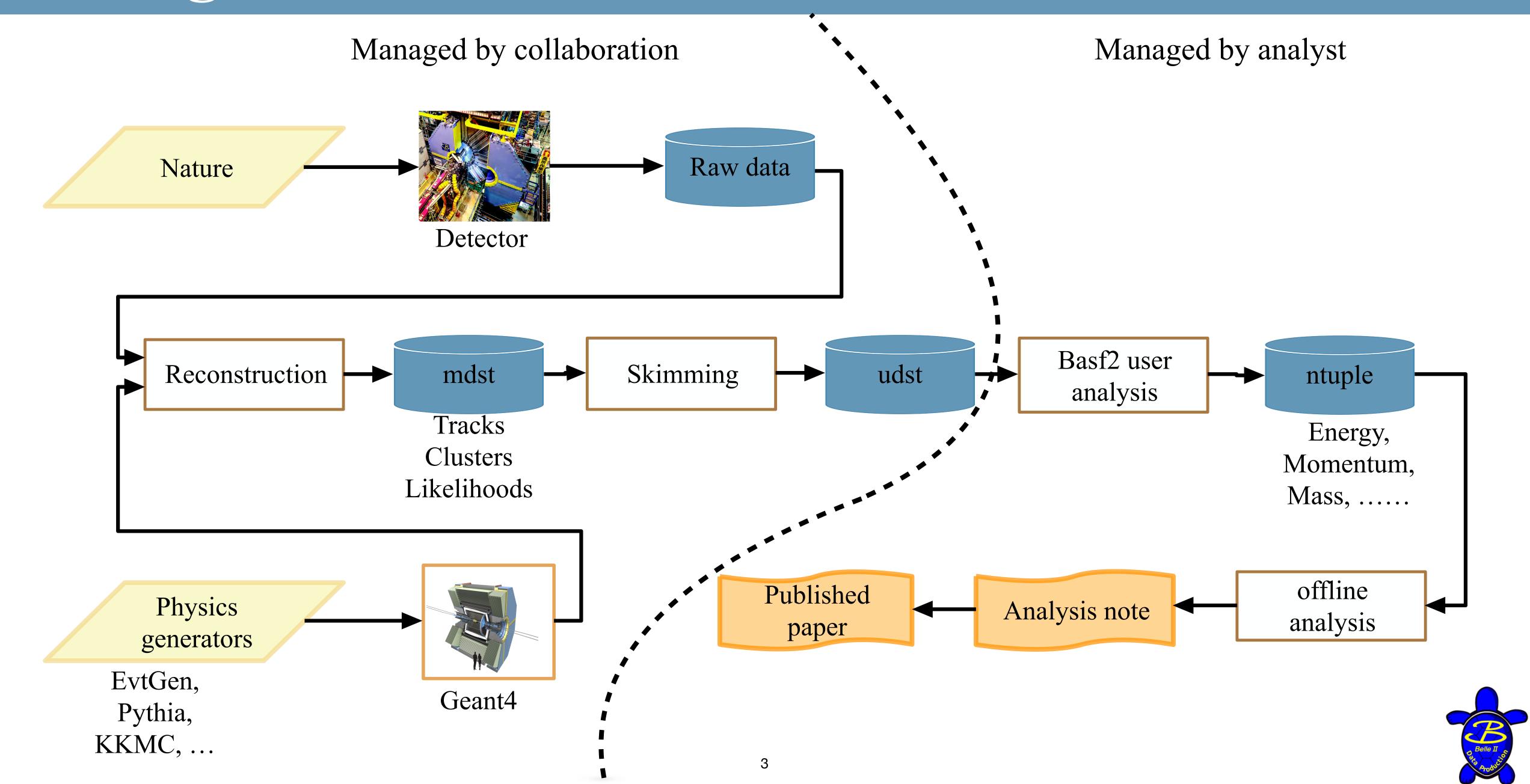
- Smooth, timely production of data and MC samples for physics analysis and other studies

• Tasks:

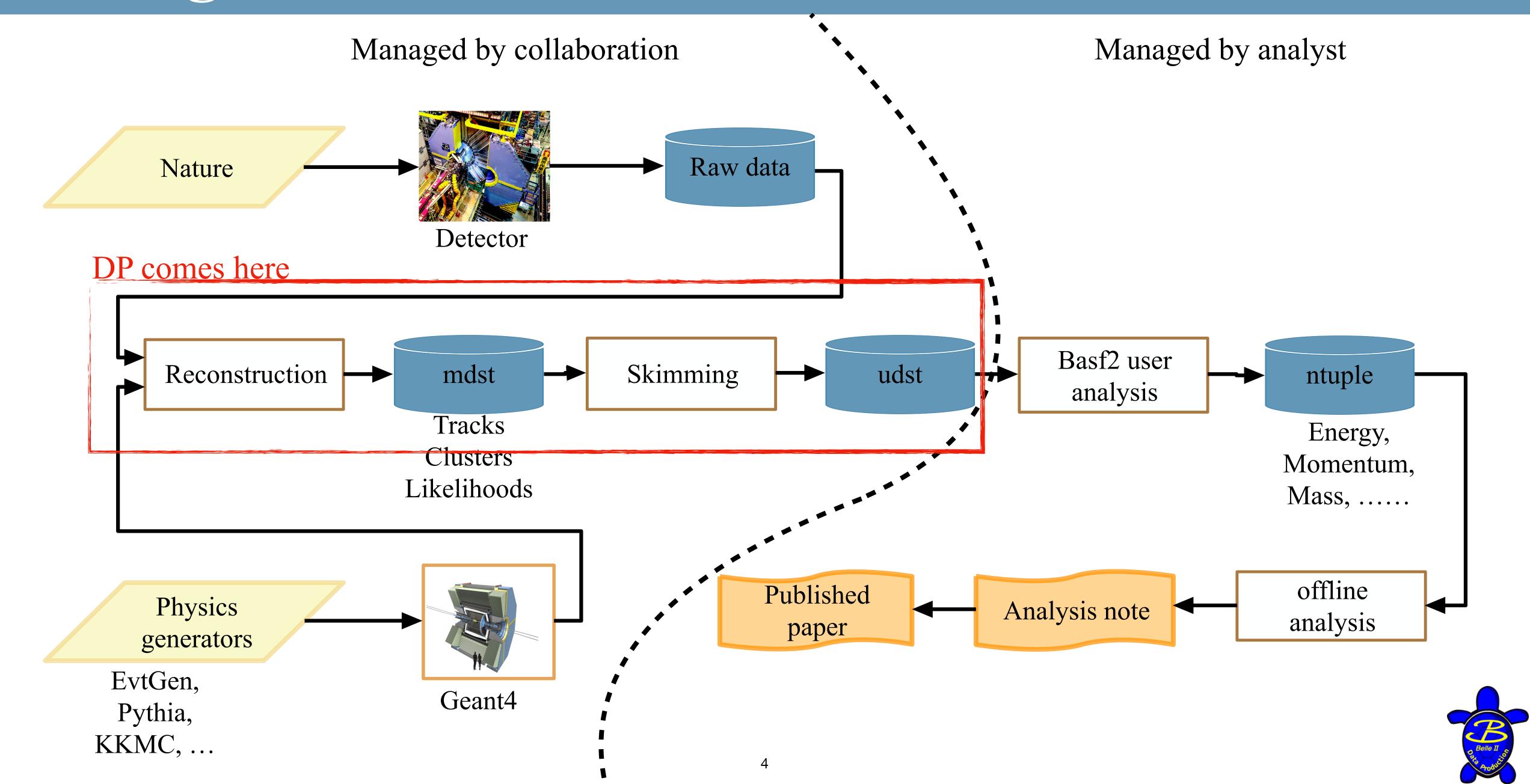
- Calibration and alignment (https://confluence.desy.de/display/BI/Data+Production+Calibration)
- RAW data (re)processing (https://confluence.desy.de/display/BI/Phase+3+data)
- MC production (https://confluence.desy.de/display/BI/Data+Production+MC12)
- Analysis skimming (https://confluence.desy.de/display/BI/Skimming+Homepage)



The Big Picture

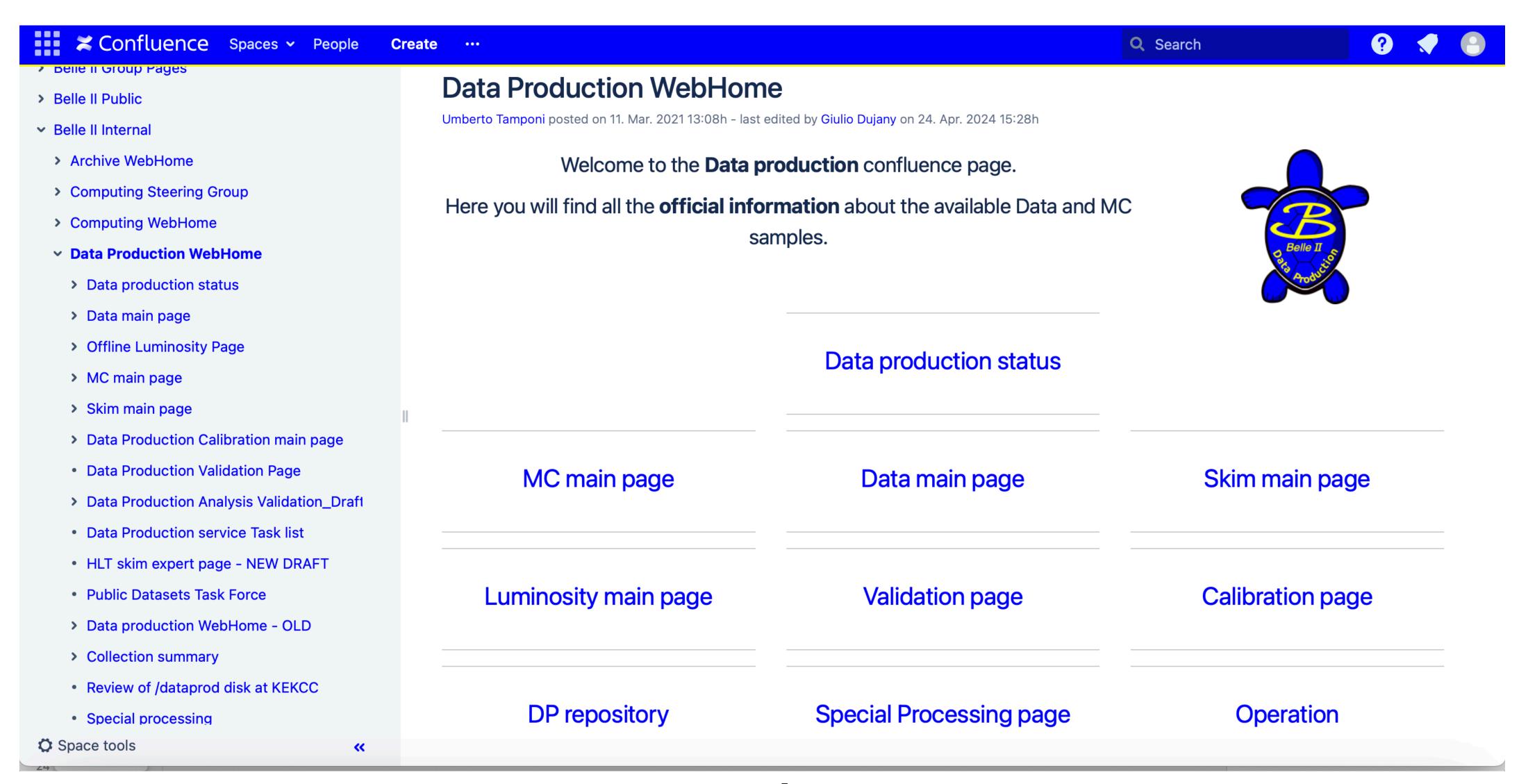


The Big Picture



Data Confluence page

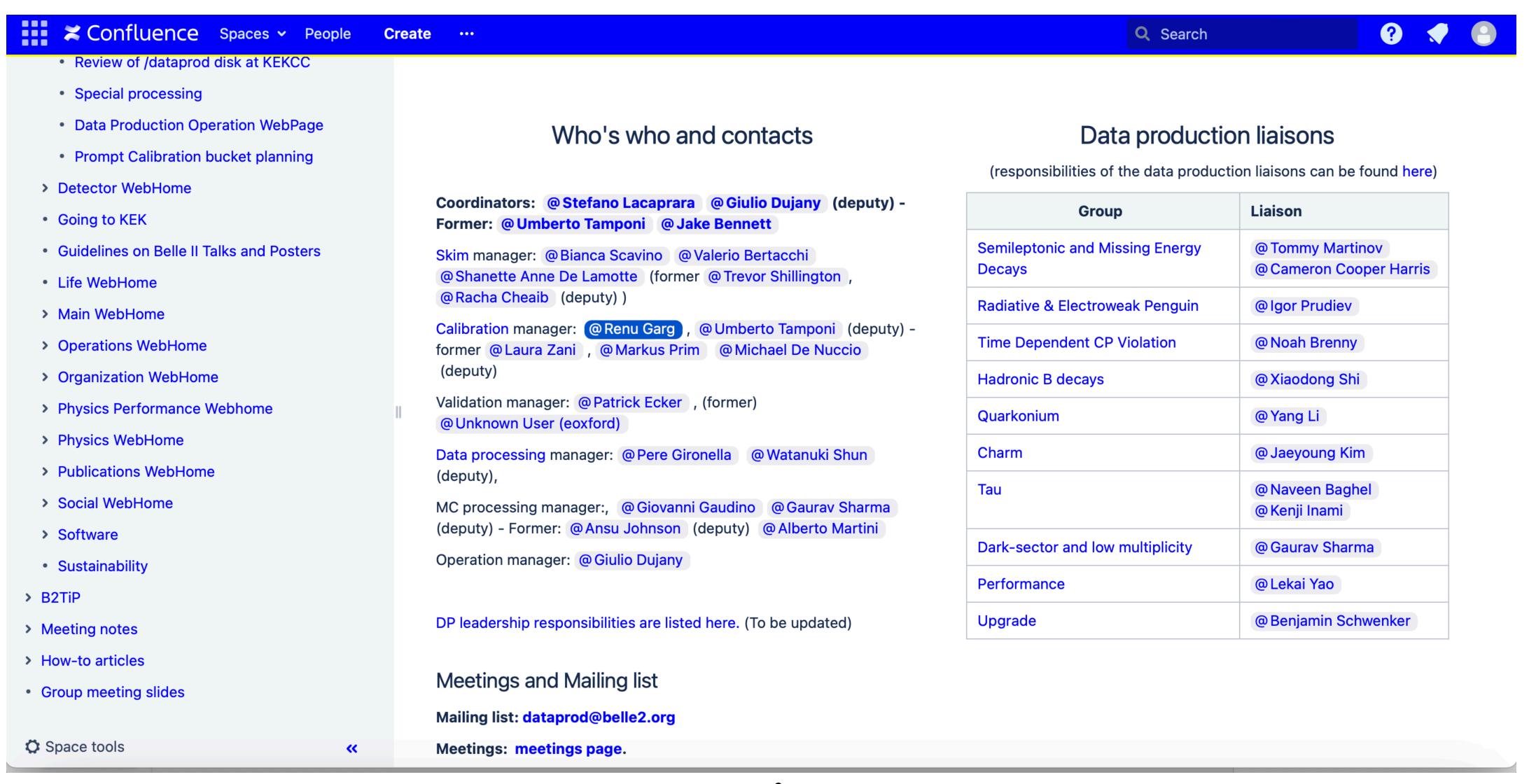
https://confluence.desy.de/display/BI/Data+Production+WebHome





Data Confluence page

https://confluence.desy.de/display/BI/Data+Production+WebHome

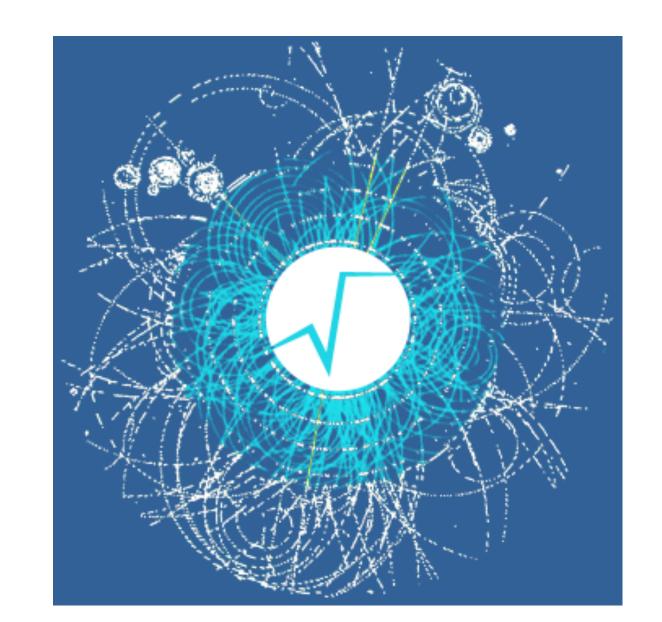




Data format

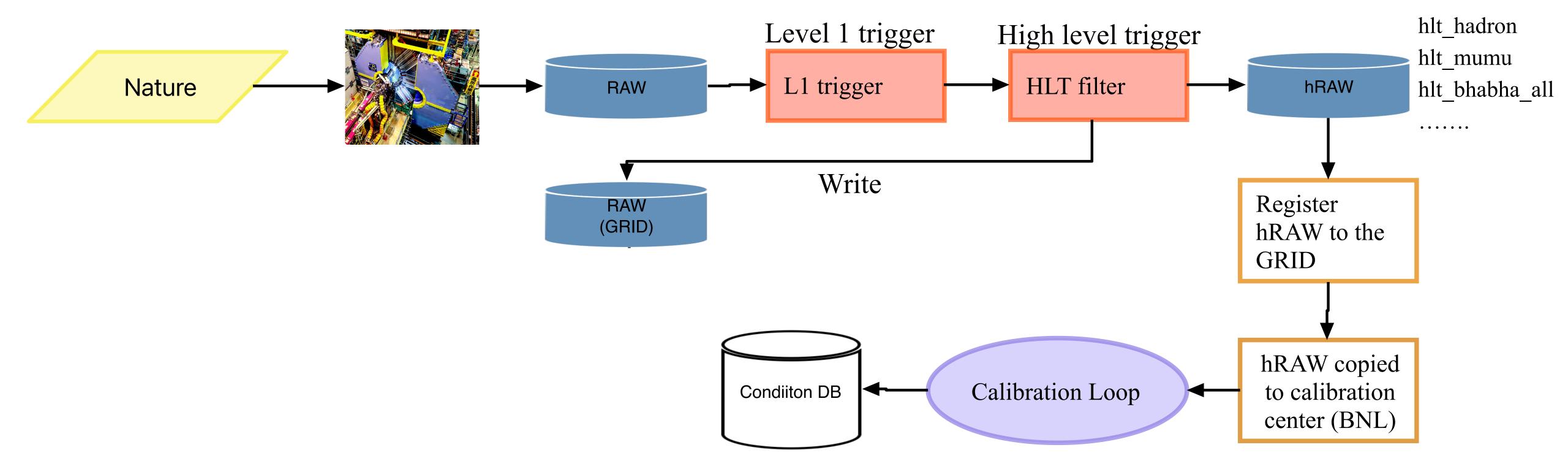
In general, Belle II output is stored in ROOT files containing various subsets of dataobjects, dbobjects, nTuples, etc

RAW:	un-processed, un-calibrated output of the detector							
hRAW:	same as RAW, but only for events passing a given HLT filter or skim							
	calibration Data Summary Table							
cDST:	- cDST contain RAW data and additional data-objects useful for calibration							
	mini Data Summary Table							
	- Controlled version of a DST.							
mDST:	- only a subset of available processed data-objects are included							
	- Flagged skim approach!							
	- Use for most analysis (see below)							
	user Data Summary Table							
D OT:	- mDST objects plus analysis objects (e.g. particleLists)							
uDST:	- produced from skims - reduce time needed to run analysis jobs							
	- Samples created only for FEI based analysis!!							





Data Flow

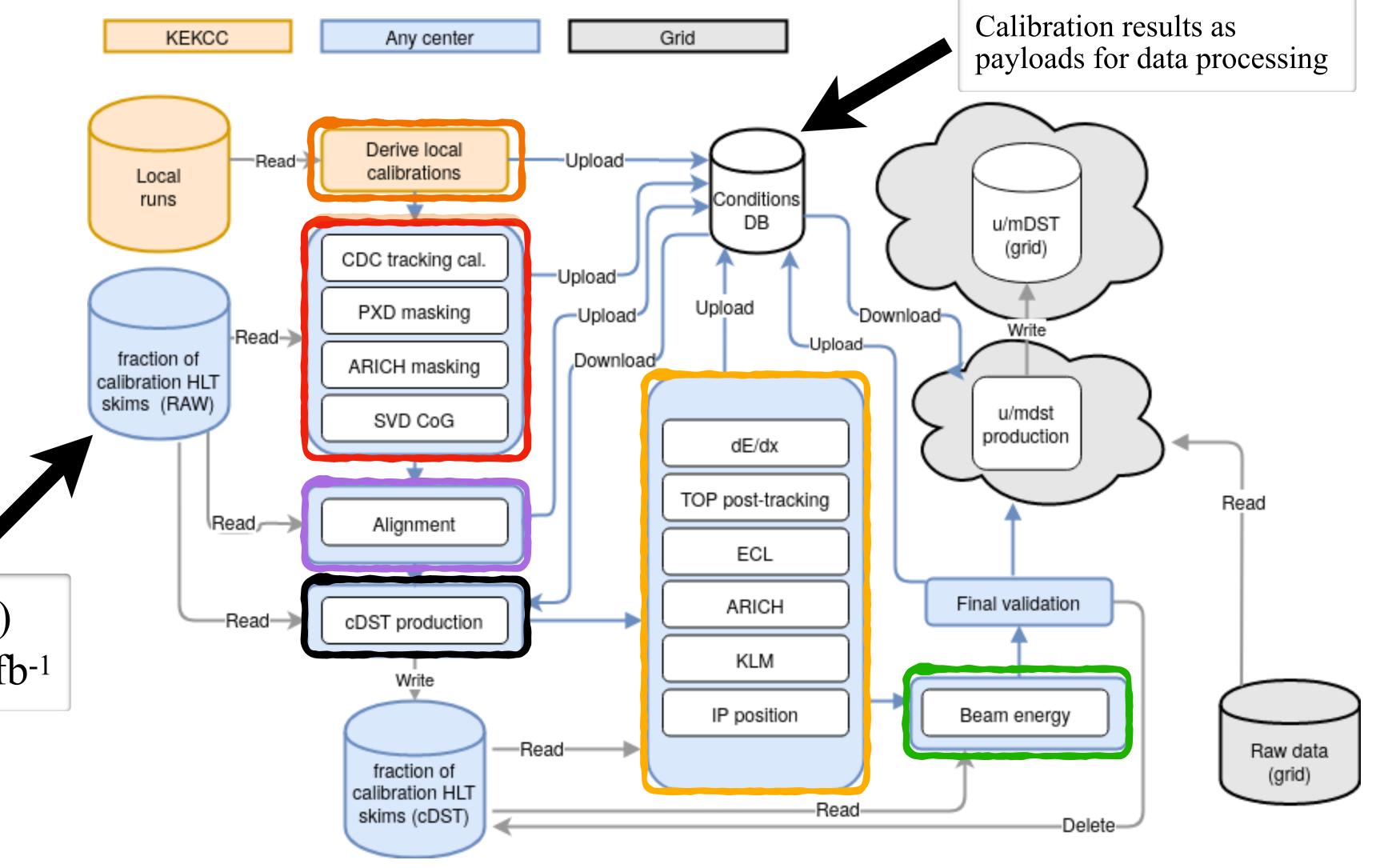




Calibration Loop

- Local calibration
- Raw data based calibration
- Alignment
- cDST production
- Post-tracking calibrations
- Analysis based calibrations

Calibration skims (hRAW)
Adaptive prescaling to 12fb-1

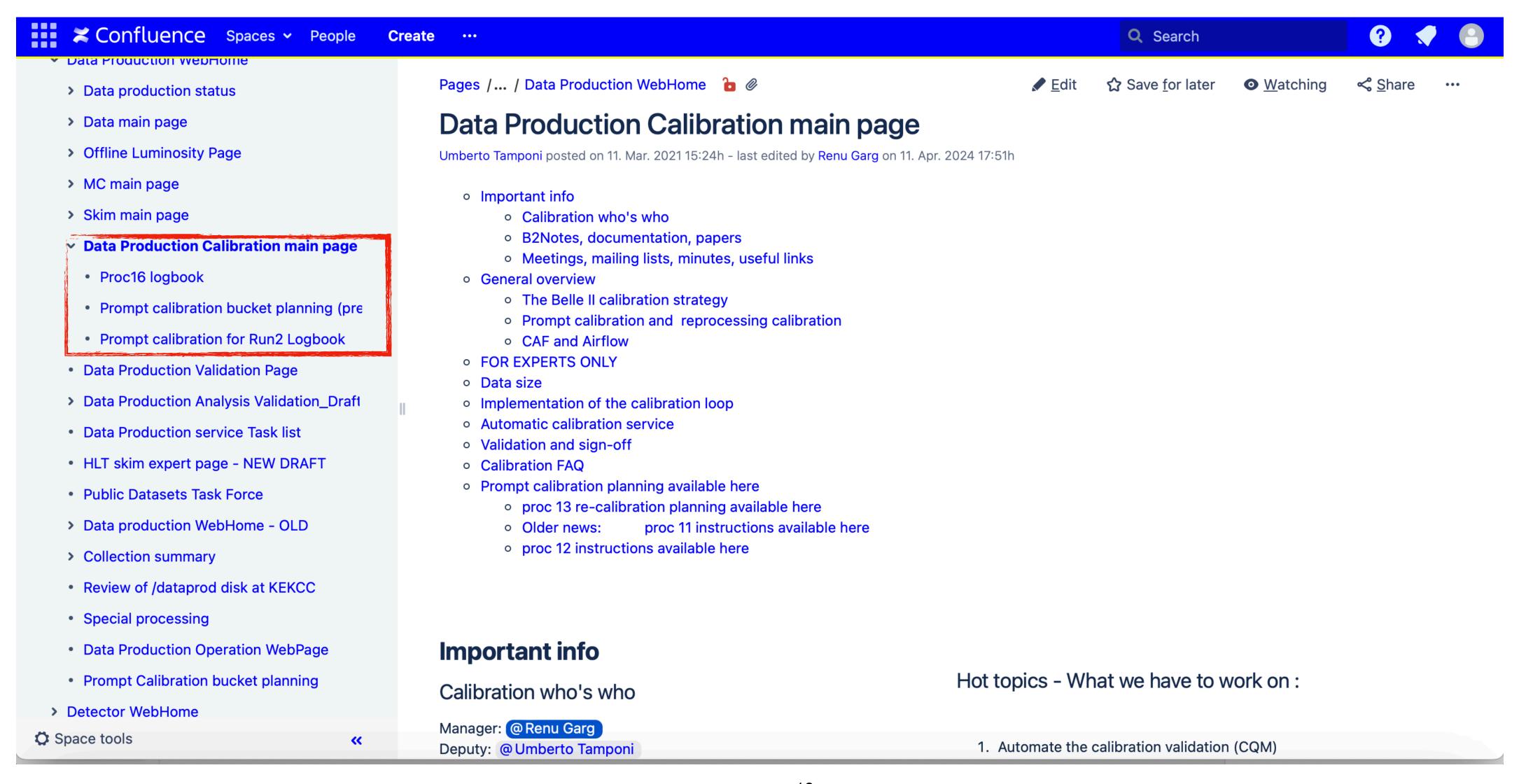


The calibration loop is fully automated



Calibration Confluence page

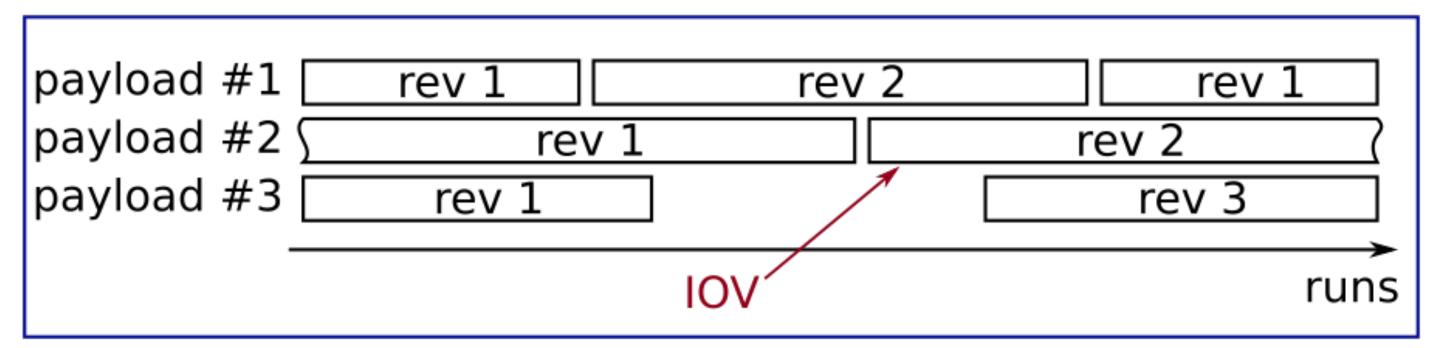
https://confluence.desy.de/display/BI/Data+Production+Calibration+main+page





Global Tag

global tag



Interval of Validity

An experiment and run interval for which a payload is valid.

An atom of condition data (e.g CDCDedxWireGain)

Payload 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

Special case:

- final_exp>=0 && final_run<0: valid for all runs in final_exp
- final_exp <0 && final_run<0:valid forever

Correct global tags are automatically selected during processing



Global Tag

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

				Questions? Write to <u>T1 CDB team</u>						
	H		obalTag oalTag: 319	Payload Types of Payload 4 items found Click on items	Global Tag for more de		n			
ID:	Name (can be partial):		Status: All	Type: All	Modifie	d by:	items	per page: 2	5 Subr	mit
	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	0	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	0	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

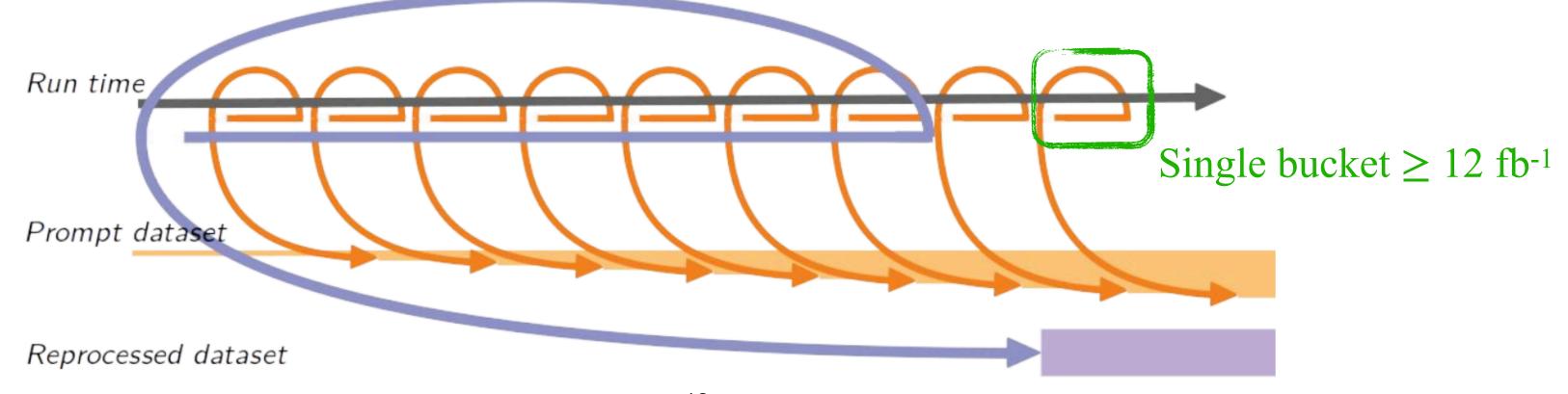
For any given data, calibration and processing happens twice:

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





(Re)Processing scheme

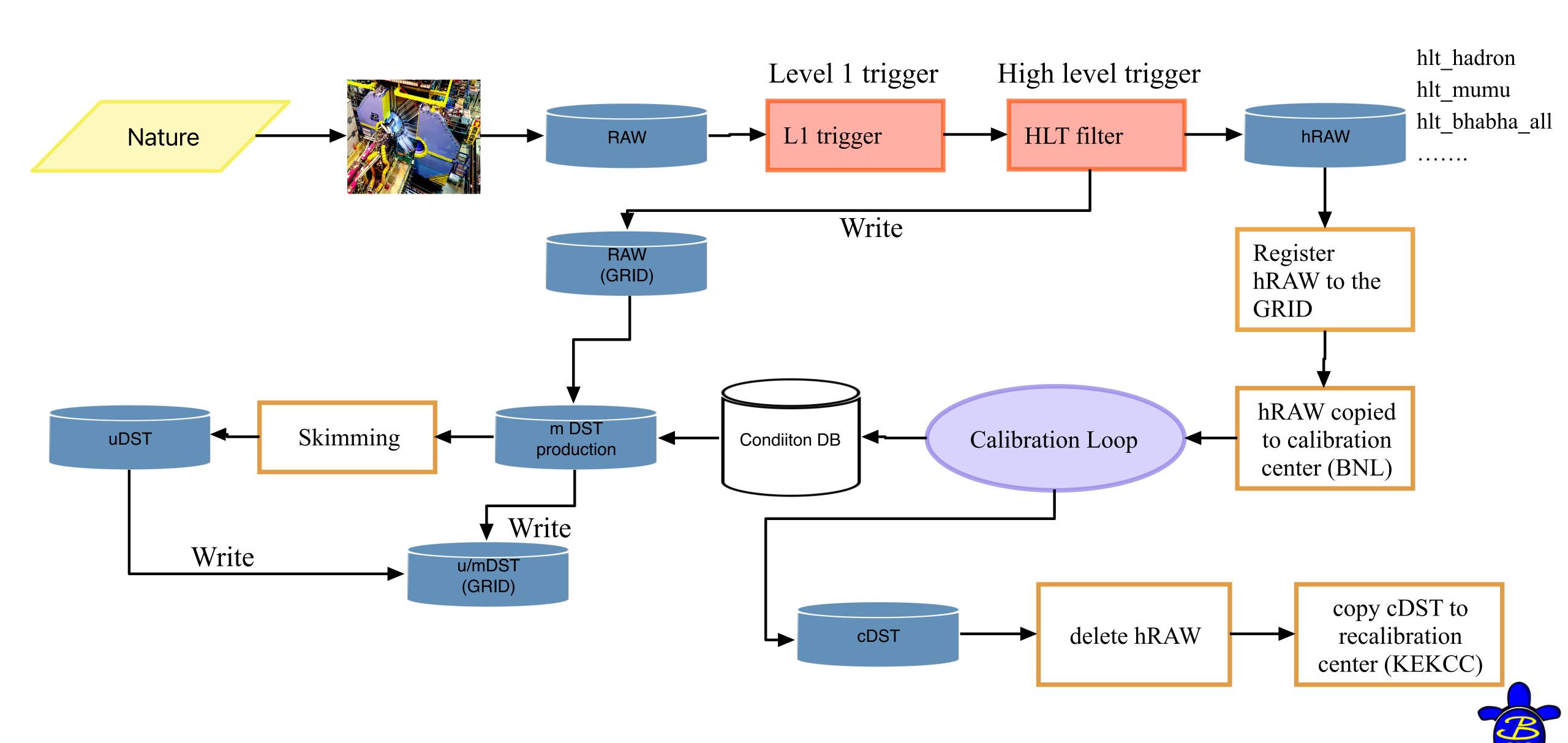
For any given data, calibration and processing happens twice:

Reprocessed dataset

Prompt processing: Official reprocessing: - Calibration (a) BNL - Recalibration @ KEKCC Minimum luminosity: 12 fb⁻¹ Uses hR Current data campaign: All calit mprovement Termind | Prompt processing: - Bucket37-39: experiment 30, 31, 32 - ~one bu - Integrated luminosity: 54.3 fb⁻¹ ler year starting - Ideally a - Will add more buckets Official processing: - Proc16: experiments 7-8, 10, 12, 14, 16-18, 20-22, 24-26 Single bucket $\geq 12 \text{ fb}^{-1}$ Prompt dataset



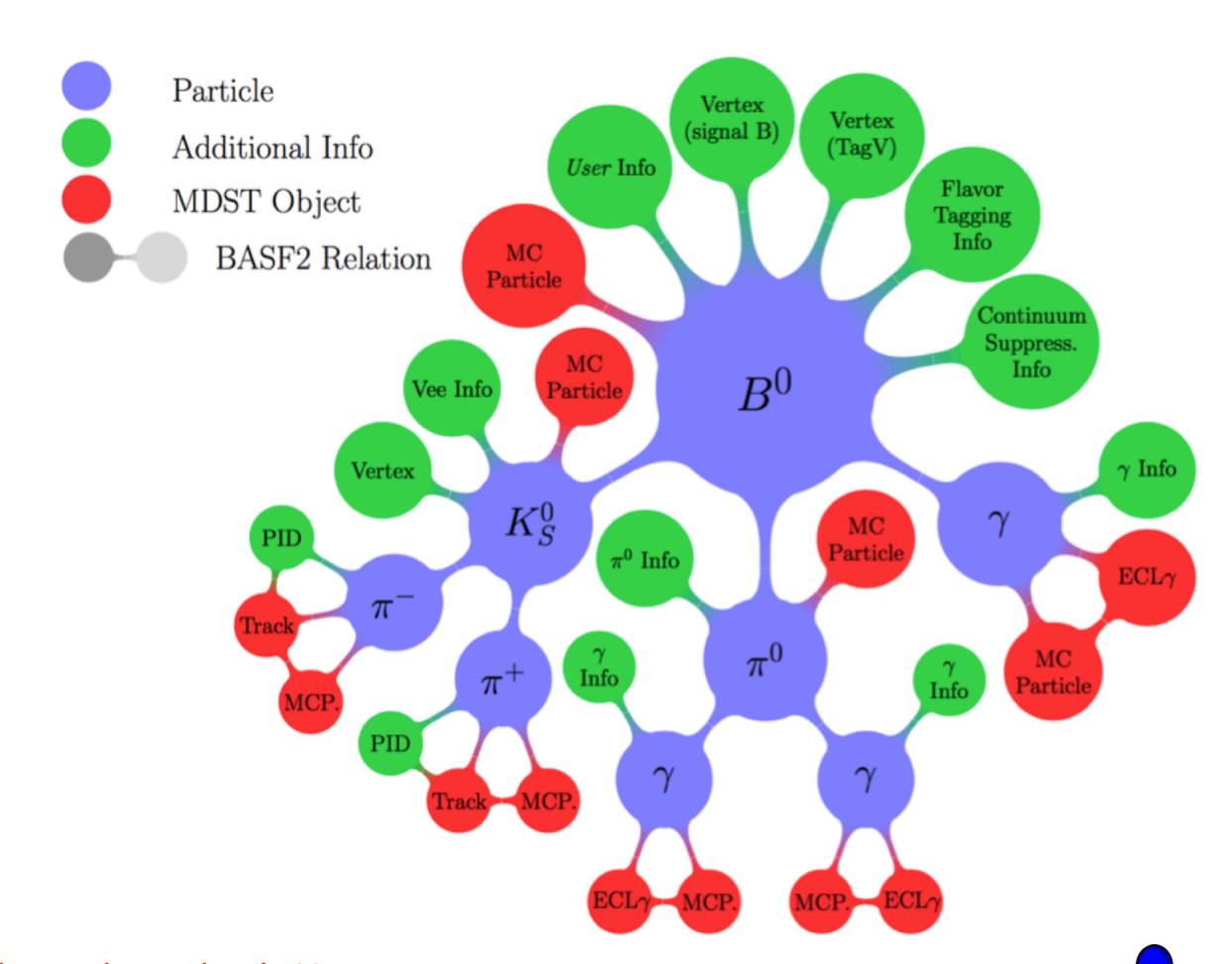
Data Flow



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:

slides, slides-1

- 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 mDST for your analysis, if not FEI based!!

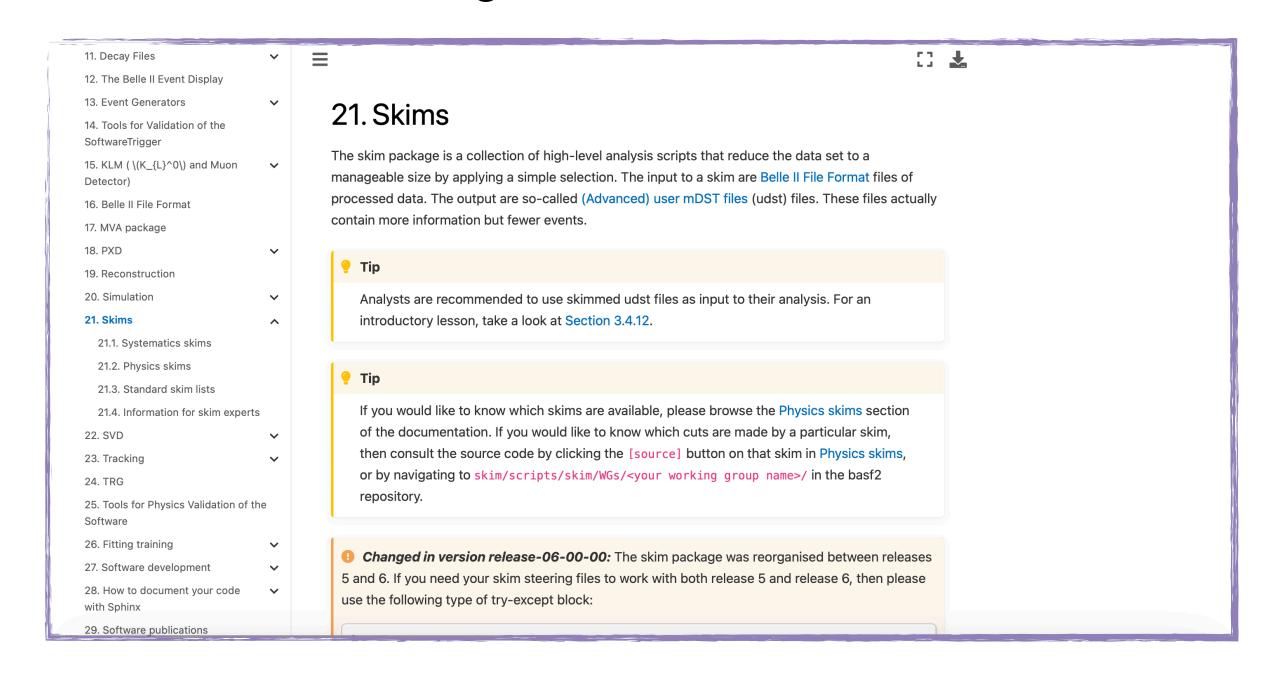


Understanding of Skimming

Two primary sources for information:

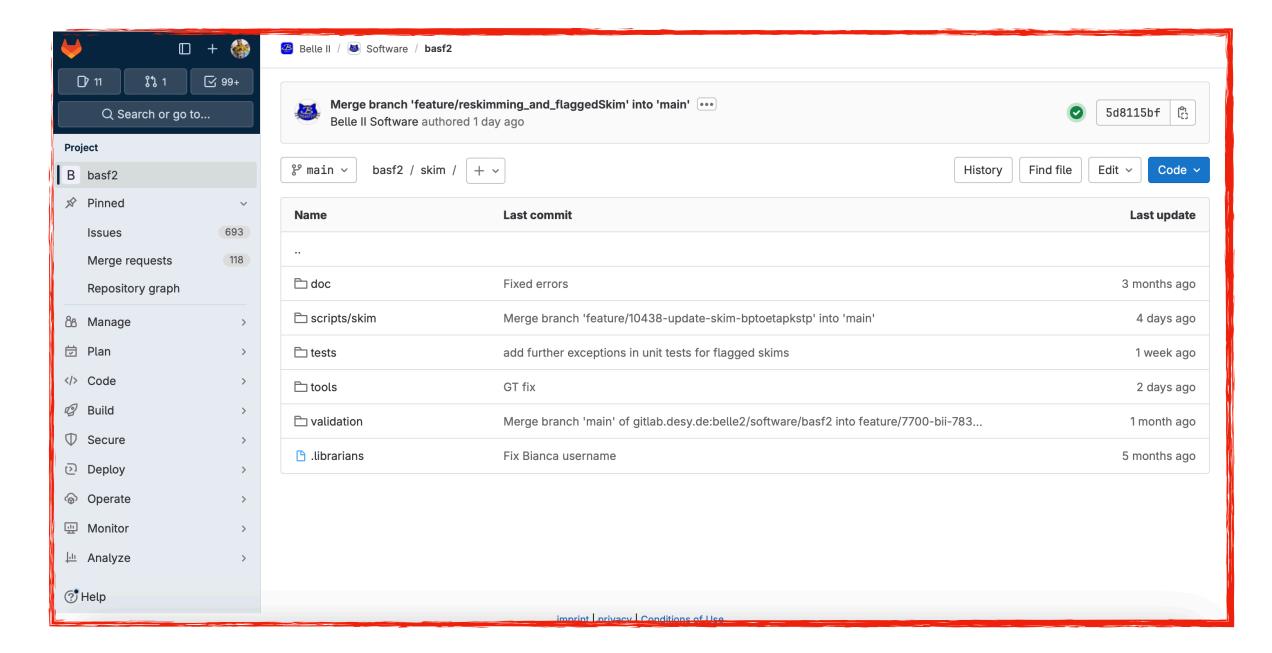
Sphinx Documentation

- 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

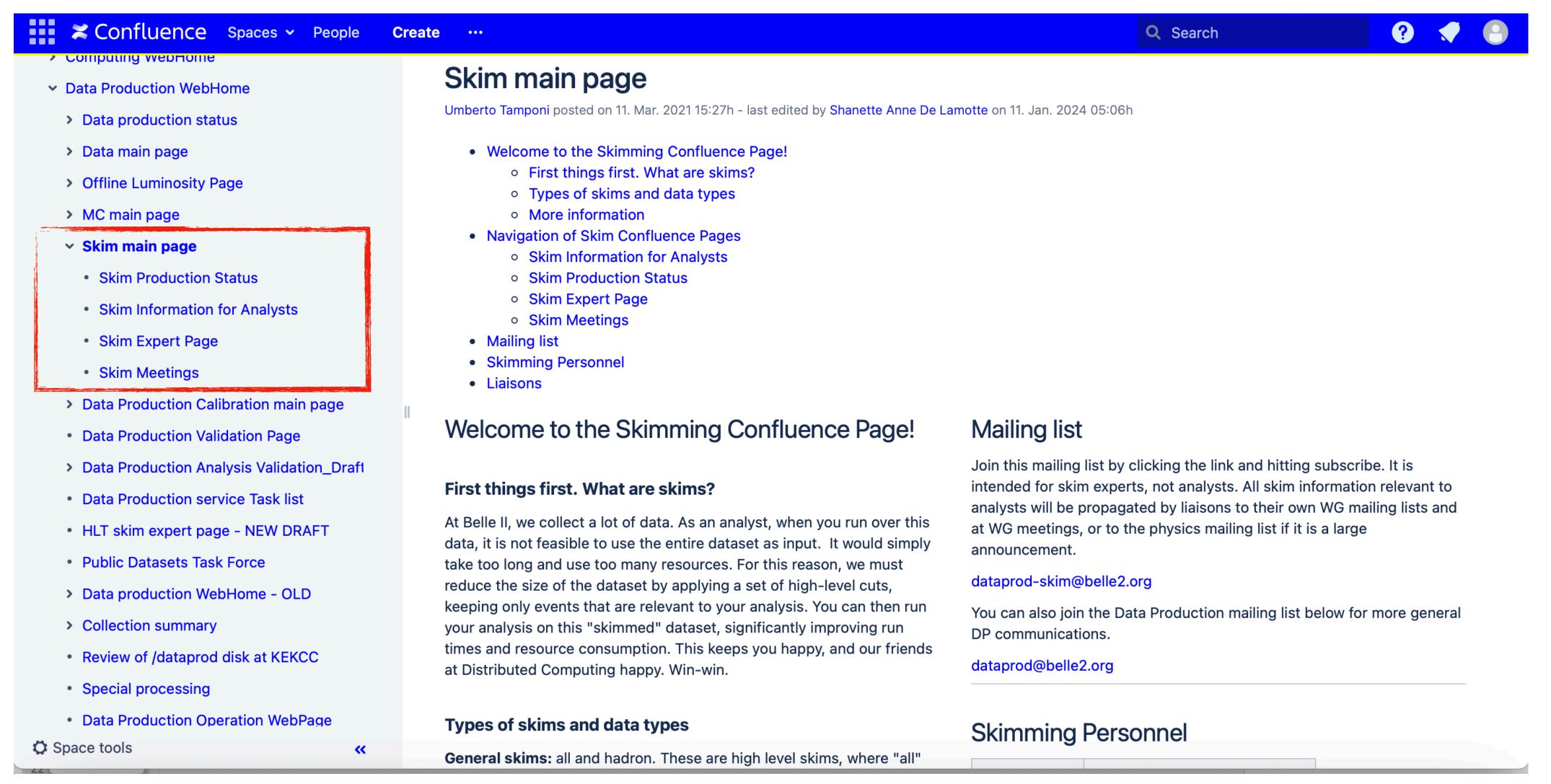
3 ways to find your skims:

- Dataset searcher on DIRAC (web version)
 - https://dirac.cc.kek.jp:8443/DIRAC/
- Dataset searcher via command line (gbasf2 environment)
 - source /cvmfs/belle.kek.jp/grid/gbasf2/pro/bashrc
 - gbasf2.belle2.org
- Collections
 - Recommended!
 - https://gbasf2.belle2.org/collectionSearcher.html
 - https://confluence.desy.de/display/BI/
 https://confluence.desy.de/display/BI/
 https://confluence.desy.de/display/BI/
 https://confluence.desy.de/display/BI/
 https://confluence.desy.de/display/BI/
 https://confluence.desy.de/display/BI/
 https://confluence.desy.de/display/BI/
 https://confluence.desy.de/display/BI/
 https://confluence.desy.de/display/BI/
 https://confluence.desy.de/display/BI/
 https://confluence.desy.de/display/BI/
 https://confluence.desy.de/display/BI/
 https://confluence.desy.de/display/BI/
 https://confluence.desy.de/display/BI/
 <a href="mailto:Skim-Information-for-Analysts-SkimCollection-for-Analysts-SkimCollection-for-Analysts-Skim-Information-for-Analysts-Skim-Information-for-Analysts-Skim-Information-for-Analysts-Skim-Information-for-Analysts-Skim-Information-for-Analysts-Skim-Information-for-Analysts-Skim-Information-for-Analysts-Skim-Information-for-Analysts-Skim-Information-for-Analysts-Skim-Information-for-Analysts-Skim-Information-for-Ana



Skim Confluence page

https://confluence.desy.de/display/BI/Skim+main+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:

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

• Generated based on central decay file: <u>DECAY_BELLE2.DEC</u>

Data production liaisons

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

Group	Liaison				
Semileptonic and Missing Energy Decays	@ Tommy Martinov @ Cameron Cooper Harris				
Radiative & Electroweak Penguin	@ Igor Prudiev				
Time Dependent CP Violation	@ Noah Brenny				
Hadronic B decays	@ Xiaodong Shi				
Quarkonium	@ Yang Li				
Charm	@ Jaeyoung Kim				
Tau	@ Naveen Baghel @ Kenji Inami				
Dark-sector and low multiplicity	@ Gaurav Sharma				
Performance	@ Lekai Yao				
Upgrade	@ Benjamin Schwenker				



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

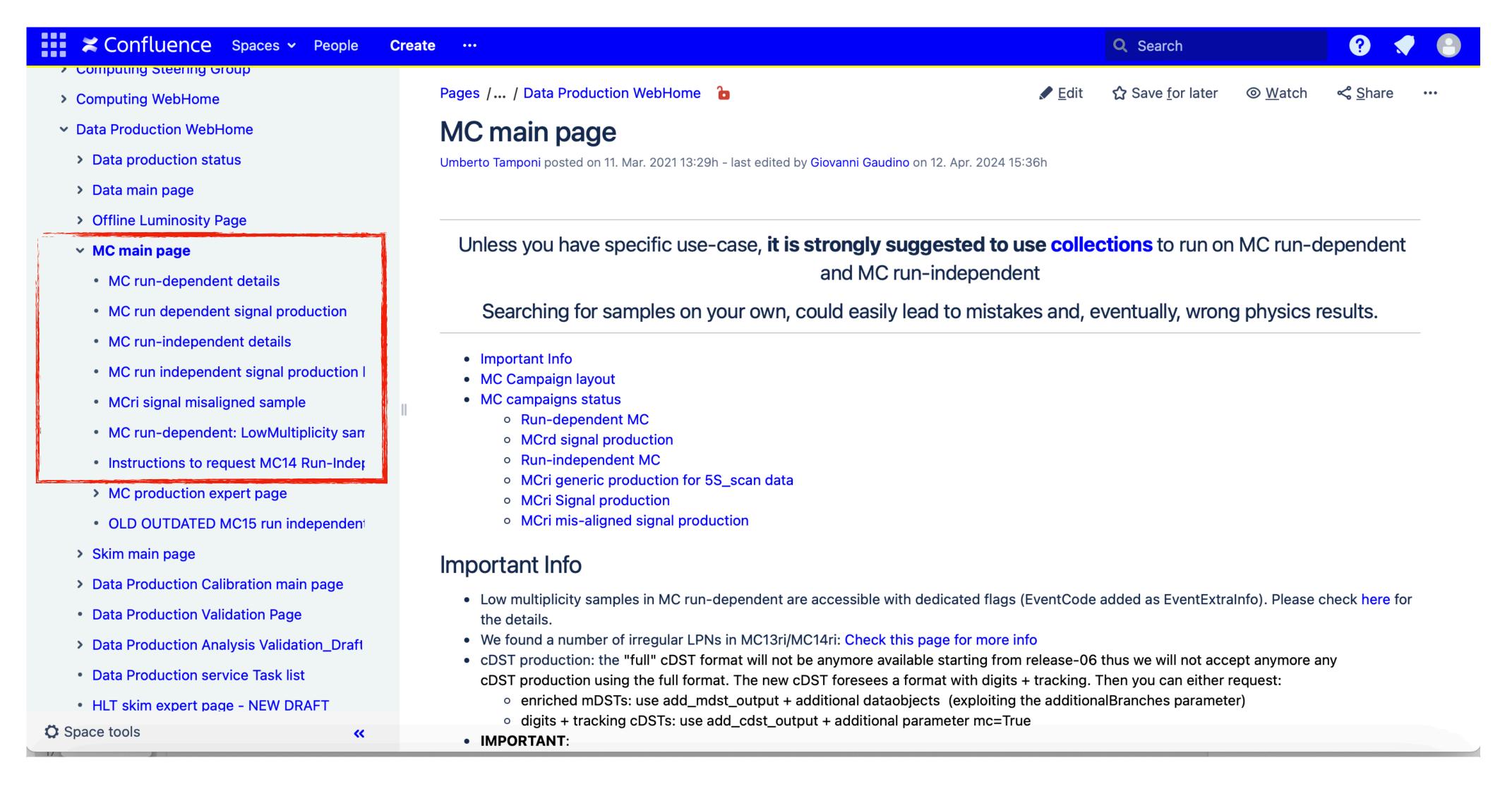
Run-dependent (RD)

- 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

https://confluence.desy.de/display/BI/MC+main+page





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/

