



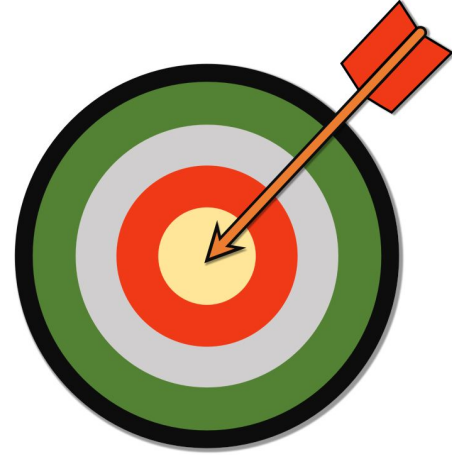
Data Production

General Overview

Trevor Shillington,
on behalf of the Data Production group

2023 Belle II Summer Workshop @ Duke University

Data Production



Primary Goal:

Central processing and simulation of official data and MC

Who are we?



DP workshop
3–9 Oct 2022
University of Roma Tre

Who are we?



Coordinators: Umberto Tamponi,
Stefano Lacaprra

Skims: Trevor Shillington, Racha
Cheaib

Calibration: Markus Prim, Michael De
Nuccio, Giulio Dujany

HLT: Gaetano de Marino

Validation: Patrick Ecker

Data processing: Watanuki Shun

MC processing: Giovanni Gaudino,
Gaurav Sharma

Global tags: Paul Laycock

DP Confluence page

The screenshot shows a Confluence page with a blue header bar containing the Confluence logo, navigation links (Spaces, People, Create), a search bar, and utility icons (help, notifications, user). The left sidebar displays a breadcrumb trail: Belle II Collaborative Services and Tools > Belle II Group Pages > Belle II Public > Belle II Internal > Archive WebHome > Computing Steering Group > Computing WebHome > **Data Production WebHome**. Under the active page, a list of sub-pages is shown, including 'Data production status', 'Data main page', 'Offline Luminosity Page', 'MC main page', 'Skim main page', 'Data Production Calibration main page', 'Data Production Validation Page', 'Data Production Analysis Validatio', 'Data Production service Task list', 'HLT skim expert page - NEW DRAI', 'Public Datasets Task Force', 'Data production WebHome - OLD', 'Collection summary', 'Review of /dataprod disk at KEKCC', 'Special processing', 'Detector WebHome', and 'Going to KEK'. The main content area features the page title 'Data Production WebHome' with a lock icon, a timestamp 'Umberto Tamponi posted on 11. Mar. 2021 13:08h - last edited by Stefano Lacaprarra on 25. May. 2023 14:48h', and a welcome message: 'Welcome to the **Data production** confluence page. Here you will find all the **official information** about the available Data and MC samples.' To the right of the text is a blue turtle logo with a yellow 'B' and 'Data Production' text. A red stamp with the text 'Bookmark this page!' is overlaid on the left side of the main content. Below the text are three columns of links, each with a horizontal line above it: 'MC main page', 'Data main page', 'Skim main page', 'Luminosity main page', 'Validation page', 'Calibration page', 'DP repository', and 'Special Processing page'. The bottom of the page shows 'Space tools' and a double-left arrow icon.

DP Confluence page

Who's who and contacts

Coordinators: @Umberto Tamponi , @Stefano Lacaprada

Skim manager: @Trevor Shillington , @Racha Cheaib (deputy)

Calibration software manager/SW liaison: @Giulio Dujany

Calibration manager: @Markus Prim , @Michael De Nuccio (deputy), former @Laura Zani

HLT skim manager: @Gaetano de Marino

Validation manager: @Patrick Ecker , (former) @Emma Oxford

Data processing manager: @Watanuki Shun ,

MC processing manager:, @Giovanni Gaudino @Gaurav Sharma (deputy) Former: @Ansu Johnson (deputy) @Alberto Martini

Global tag manager: @Paul Laycock

DP leadership responsibilities are listed here.

Meetings and Mailing list

Mailing list: dataprod@belle2.org

Meetings: [meetings page](#).

Minutes: (2022) <https://hackmd.io/dbskL9vDQjeQ1PXuu-Zzog>
(2023) <https://hackmd.io/vGeJNbmQSV6F01cY3bVZ0w>

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	@Moritz Bauer [was @Mario Merola]
Radiative & Electroweak Penguin	@Soumen Halder , @Filippo Dattola
Time Dependent CP Violation	@Yongqing Chen
Hadronic B to Charmless	@Emilie Bertholet
Hadronic B to Charm	@Yi Zhang
Bottomonium	@Unknown User (justing)
Charmonium	@Yang Li
Charm	@Michel Bertemes [was @Emma Oxford]
Tau	@Swagato Banerjee
Dark-sector and low multiplicity	@Giacomo De Pietro

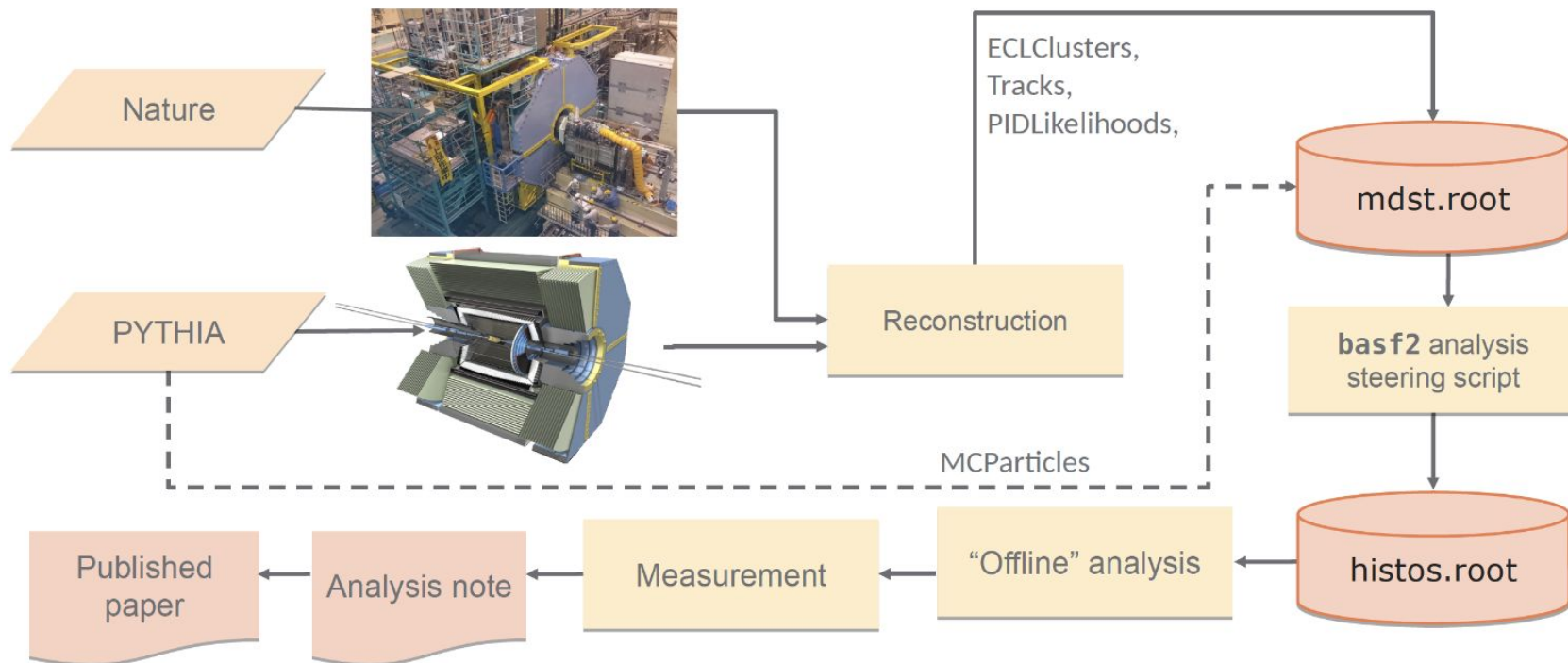
Service tasks

Help Data Production in limited and well defined task to earn you authorship in BelleII.

- [Data Production service Task list](#)

Know your WG liaison!

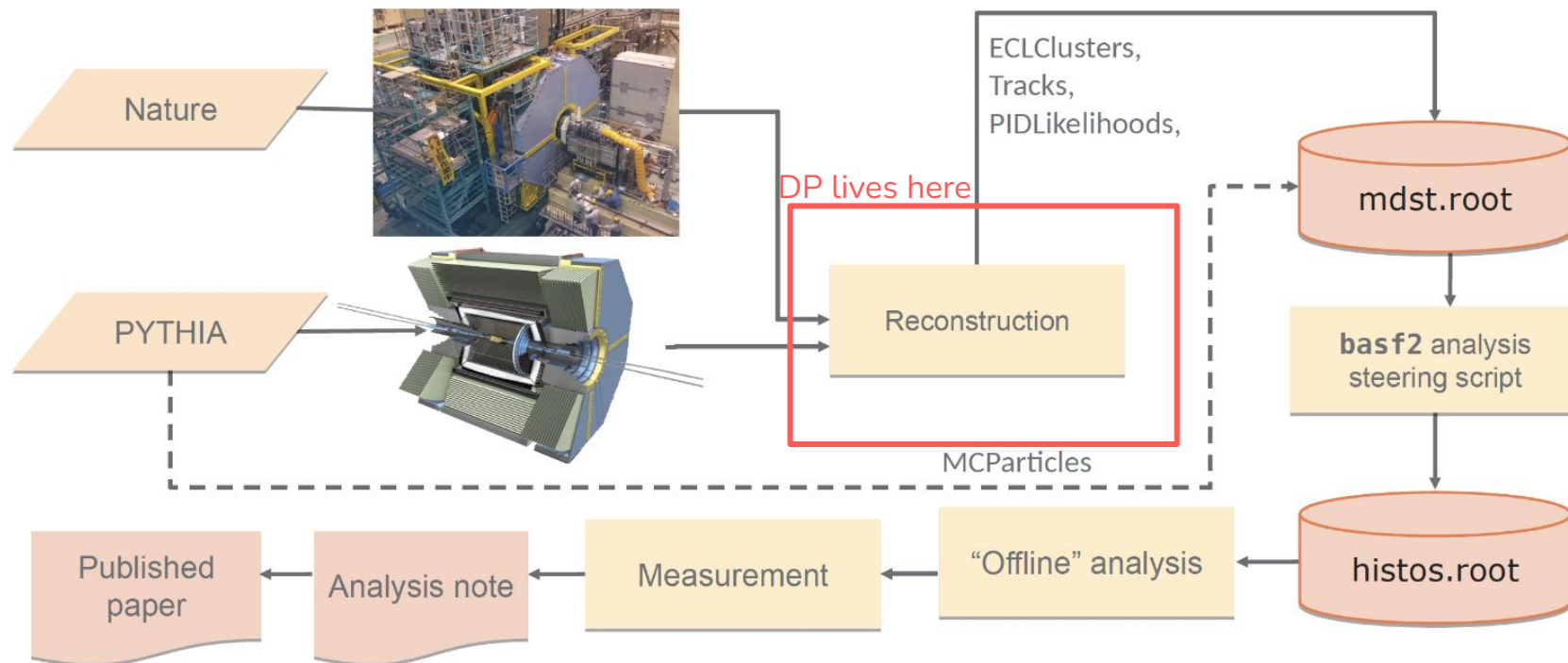
The big picture



*shamelessly stolen borrowed from Sam Cunliffe's talk "Introduction to the analysis package" - Belle II SKW, 15.06.2018

*shamelessly stolen again borrowed from Jake Bennett's DP talk at 2022 Belle II Summer Workshop

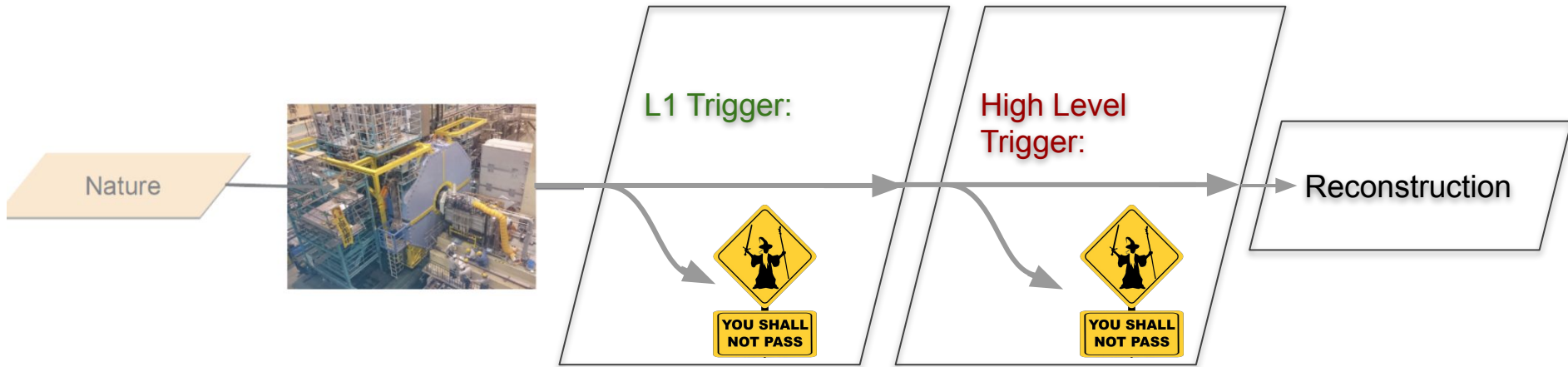
The big picture



*shamelessly stolen borrowed from Sam Cunliffe's talk "Introduction to the analysis package" - Belle II SKW, 15.06.2018

*shamelessly stolen again borrowed from Jake Bennett's DP talk at 2022 Belle II Summer Workshop

Overview: Data flow



Level 1 Trigger (TRG or L1) looks at low res “live stream” from CDC, ECL, KLM

If decision to keep event is made by **L1**, all detectors transmit readout data to event builder and **High Level Trigger (HLT)** units

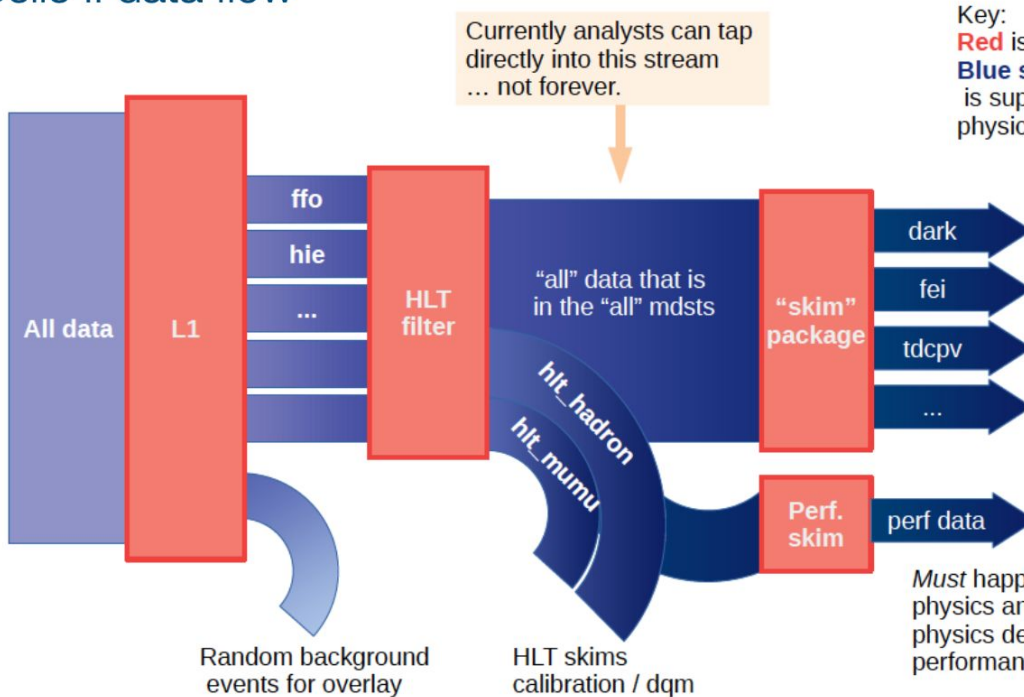
HLT (computing cluster w/ ~10k cores) reconstructs full subdetector data per event, then classifies it for storage or deletion (**60% reduction**)

Prescaling: trigger only some fraction of given event (e.g. we want *some* Bhabha events but not *all* Bhabha events)

Overview: Data flow

Image credit: S. Cunliffe

Belle II data flow



Key:
Red is a filter
Blue scale from light to dark
is supposed to indicate more
physics-relevant data

Notes:

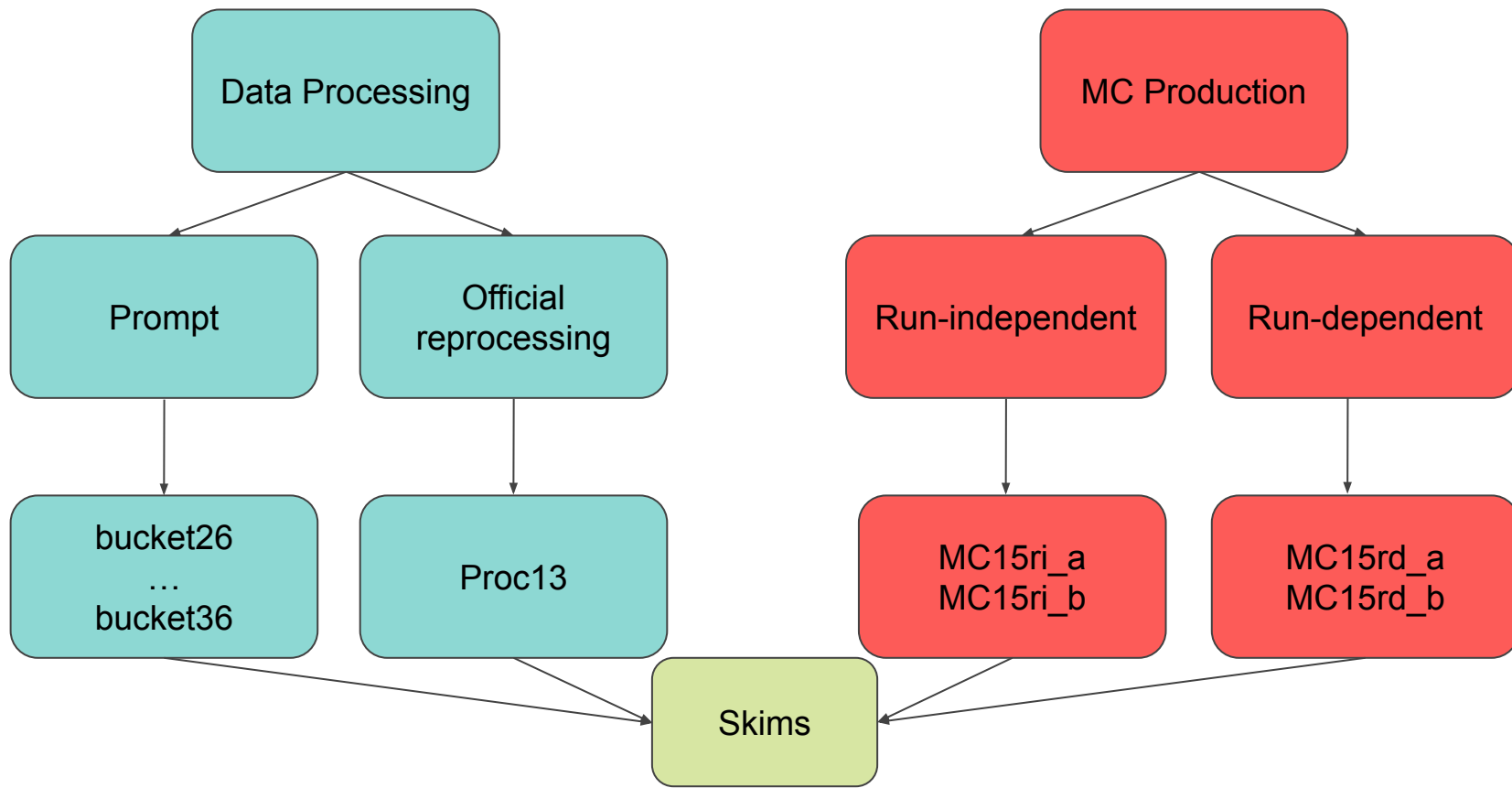
- **hlt_hadron is about 10% of "all" data**
- **About 50% of analysis skims run on hlt_hadron, the rest run on "all" data**

Must happen faster than physics analysis skims: physics depends on performance analyses

* **hlt_hadron** = at least 3 "good" tracks ($pt > 0.2$, $d_0 < 2$, $abs(z_0) < 4$) and NOT Bhabha-like



Overview: Data flow



DP Jargon: Data types

RAW: un-processed, un-calibrated output of the detector

hRAW: same as RAW, but only for events passing a given HLT filter or skim

- Use RAW data to reconstruct tracks, showers, etc. to get a **data summary table (DST)**

cDST: calibration data summary table

- cDST contain RAW data and additional dataobjects useful for calibration

mDST: mini data summary table

- Controlled version of a DST.
- Curated list of post-reconstruction dataobjects for analysis use.
- MC and data campaigns output mdst.

uDST: user data summary table

- mDST objects plus analysis objects (e.g. particleLists)
- Analysis skims output uDST
- **Smallest file sizes** → **reduces runtime for analysis jobs (unclogs the grid)**
- **USE THESE!!**



→ **Just these are relevant to analysts**



(Re)Processing Data

Experiment: A longer period of experimental data taking. Numbered sequentially.

- The most recent is experiment 27*

run[‡]: A period of uninterrupted data taking (from minutes to hours).

- Conditions[†] can change between runs
- Hundreds to thousands of runs per experiment

Event: One readout of the detector.

- Every single event is **uniquely** identified by **(exp, run, evt)**

generalSkimName: “all” or “hadron”, indicating whether the data is processed on “all” HLT events or “hlt_hadron” skimmed HLT events

[†] **Conditions:** Calibrations and other data which might vary per run but are not part of the event

[‡]A run/Run can have various meanings: <https://confluence.desy.de/display/BI/Main+Glossary#MainGlossary-R>

* <https://confluence.desy.de/display/BI/Experiment+numbering>



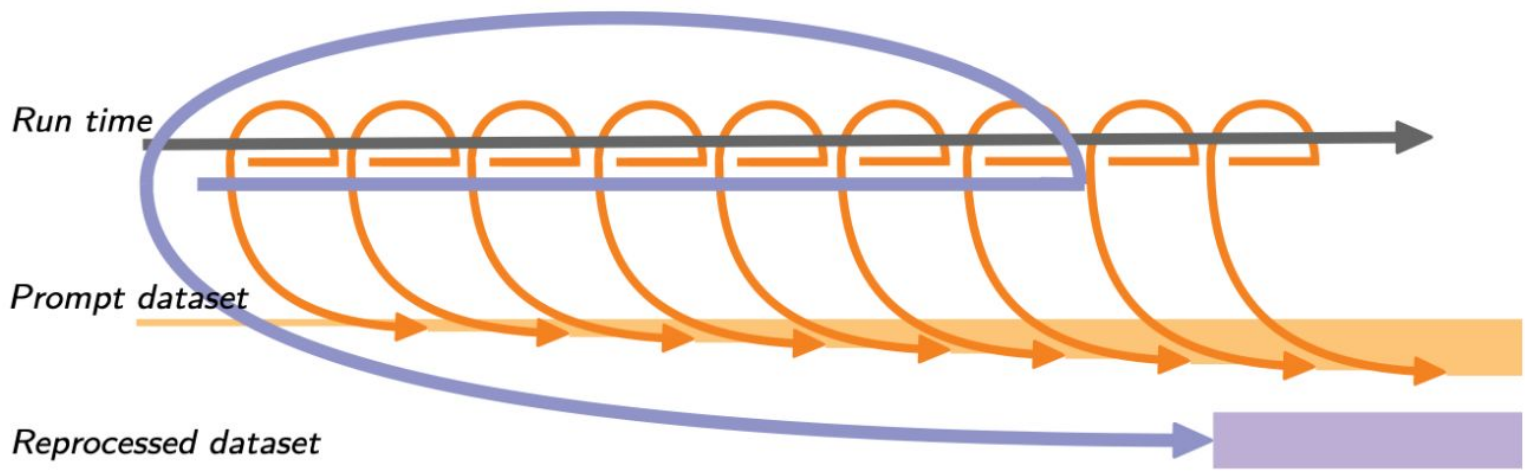


(Re)Processing Data

For any given data, calibration and processing happens **twice**:

Prompt processing: ~weekly during data taking → “buckets” of runs with 9-20 fb⁻¹ → **bucketXX**

Official reprocessing: ~yearly to make final changes and incorporate calibrations that require more data → **procXX**

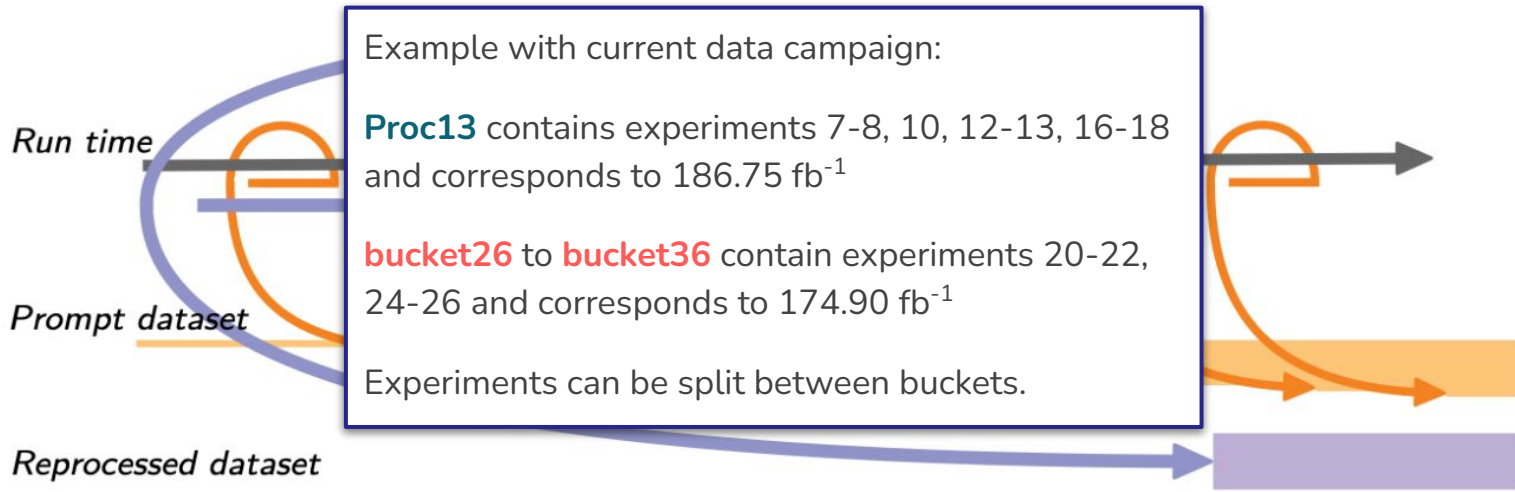


(Re)Processing Data

For any given data, calibration and processing happens **twice**:

Prompt processing: ~weekly during data taking → “buckets” of runs with 9-20 fb⁻¹ → **bucketXX**

Official reprocessing: ~yearly to make final changes and incorporate calibrations that require more data → **procXX**

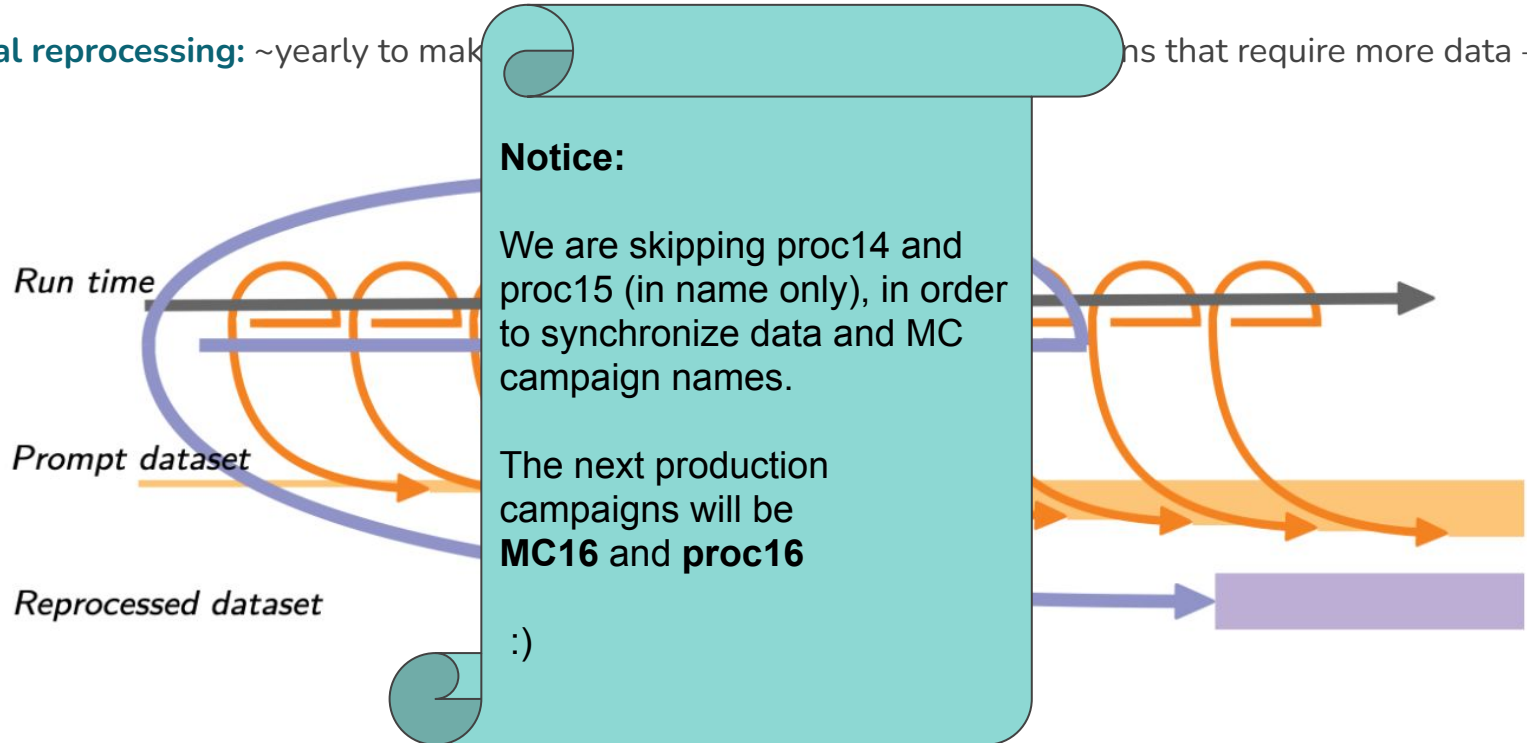


(Re)Processing Data

For any given data, calibration and processing happens **twice**:

Prompt processing: ~weekly during data taking → “buckets” of runs with 9-20 fb⁻¹ → **bucketXX**

Official reprocessing: ~yearly to make campaigns that require more data → **procXX**



- > Belle II Public
- ▼ Belle II Internal
 - > Archive WebHome
 - > Computing Steering Group
 - > Computing WebHome
 - ▼ Data Production WebHome
 - **Data production status**
 - > Data main page
 - > Offline Luminosity Page
 - > MC main page
 - > Skim main page
 - Data Production Calibration main p
 - Data Production Validation Page
 - > Data Production Analysis Validatio
 - Data Production service Task list
 - HLT skim expert page - NEW DRAI
 - Public Datasets Task Force
 - > Data production WebHome - OLD
 - > **Collection summary**
 - Review of /dataproduct disk at KEKCC
 - Special processing
 - > Detector WebHome
 - Going to KEK

Pages /... / Data Production WebHome

Edit Save for later Watching Share

Data production status

Good resources!

Umberto Tamponi posted on 29. Oct. 2021 10:44m - last edited by Stefano Lacaprara on 15. Mar. 2023 09:59h

Available processing

- This page has a snapshot of what is available in Data and MC (both run independent and dependent) processing. There is a table for each processing campaign (latest on top).
- To access data or MC, please use collections as described on [Collection summary](#)
- The luminosity reported is the offline one, please refer to [Offline Luminosity Page](#)
- Detailed information about data processing are [Data main page](#) for MC are [MC main page](#)
- link to Physic Performance recommendation and systematics
- color legend: **GREEN: the production is ready.** **YELLOW: processing is still running**
- Available processing
 - Release-06: proc13 - MC15
 - Release-05: proc12 - MC14

Release-06: proc13 - MC15

Exp	Offline. Luminosity (/fb)	Data (hadron)	Data (all)	MC run dependent	MC run independent	No
7	4S: 0.510	proc13	proc13	Done	MC15ri	
8	4S: 4.459 4S_offres: 0.813 4S_scan: 0.038	collections: Data main page please use merged collections	collections: Data main page		collections: MC main page#Run-independentMC	Rui ha
10	4S: 3.635					
12	4S: 54.388 4S_offres: 8.716					
14	4S: 16.385					



Calibration

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

A full calibration loop is divided into 5 steps, each one depending on the previous ones:

1. **Local calibrations**
2. **Raw-data based calibrations**
3. **Alignment**
4. **Post-tracking calibrations**
5. **Analysis-based calibrations**

Calibration

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

A full calibration loop is divided into 5 steps, each one depending on the previous ones:

- 1. Local calibrations**
2. Raw-data based calibrations
3. Alignment
4. Post-tracking calibrations
5. Analysis-based calibrations

- Derived by local runs or DQM.
- Examples:
 - TOP laser calibrations
 - SVD noise calibration
 - TOP channel masking

Calibration

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

A full calibration loop is divided into 5 steps, each one depending on the previous ones:

1. Local calibrations
2. **Raw-data based calibrations**
3. Alignment
4. Post-tracking calibrations
5. Analysis-based calibrations

- Must run on raw collision data
- Do not require good tracks
- Examples:
 - Channel masking
 - CDC tracking calibration

Calibration

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

A full calibration loop is divided into 5 steps, each one depending on the previous ones:

1. Local calibrations
2. Raw-data based calibrations
3. **Alignment**
4. Post-tracking calibrations
5. Analysis-based calibrations

- Requires raw collision data
- Example:
 - Corrections to the position of tracking detector sensors

Calibration

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

A full calibration loop is divided into 5 steps, each one depending on the previous ones:


1. Local calibrations
2. Raw-data based calibrations
3. Alignment
4. **Post-tracking calibrations**
5. Analysis-based calibrations

- Require good tracks
- Run on cDST
- Examples:
 - CDC dE/dx
 - IP position

Calibration

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

A full calibration loop is divided into 5 steps, each one depending on the previous ones:

1. Local calibrations
 2. Raw-data based calibrations
 3. Alignment
 4. Post-tracking calibrations
 5. **Analysis-based calibrations**
- 

- Rely on high quality data
- Example:
 - Beam energy



Calibration

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

A full calibration loop is divided into 5 steps, each one depending on the previous ones:

1. Local calibrations
2. Raw-data based calibrations
3. Alignment
4. Post-tracking calibrations
5. Analysis-based calibrations

Notes:

- Calibration runs twice
 - Prompt calibration
 - Reprocessing calibration
- Calibrations are run automatically via Airflow
- Takes about 9 fb^{-1} of data to re-derive calibrations
 - Hence, buckets are at least 9 fb^{-1}
- A full calibration takes ~ 15 days using 1000 cores

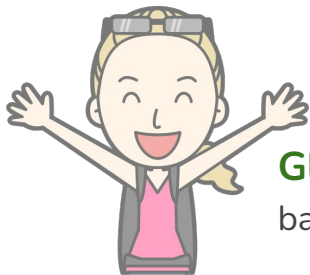
A quick note on Global Tags

Conditions Database: place where we store additional data, like detector configuration or calibration constants

Global tag: immutable collection of payloads for a certain dataset

↳ **Payloads:** one atom of conditions data (a file)

↳ **IOV:** “Interval of Validity”, the experiment and run interval for which the payload is valid.



GlobalTag replay: Correct global tags are automatically selected during processing, based on what was used to create the input file.

Note: specifying a global tag is usually only done in expert settings

A quick note on Global Tags

Conditions Database: <https://cdbweb.sdcc.bnl.gov>

Can be useful!

Questions? Write to [T1 CDB team](#)

Home GlobalTag Payload Types of Payload Global Tag Comparison

GlobalTag: 523 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	Distinct Payl
payloads_for_exp1004_from_proc13	2957	●	Payloads for exp. 1004 from proc13	TESTING	DEV	06/15/2023 10:30 p.m.	gdepietro	1	1
mc_production_sproc4_tracking	2949	●	GT for the sproc4 MCrd production, requested by tracking group	TESTING	DEV	05/16/2023 9:02 a.m.	gaudino	1	1
svd_proc16	2946	●	Update of SVD payloads for proc16	OPEN	DEV	05/09/2023 9:08 a.m.	gdujany	81	11
data_reprocessing_s-proc4	2920	●	Global tag including calibration payloads from s-proc4 and baseline payloads.	VALIDATED	DEV	03/23/2023 10:21 a.m.	mapr	3224	294
data_reprocessing_s-proc4_broken2	2919	●	Final staging tag to collect and test all calibrations in sproc4.	OPEN	DEV	03/22/2023 4:10 p.m.	mapr	3048	66
AIRFLOW_final_staging_calibration_sproc4_snapshot20230322	2918	●	Final staging tag to collect and test all calibrations in sproc4.	OPEN	DEV	03/22/2023 1:48 p.m.	ddossett	3049	66
data_reprocessing_s-proc4_broken	2917	●	Baseline globaltag for the s-proc4 campaign to recalibrate bucket36 with prerelease-07-00-00c. Includes in order of priority patch_main_release-07_noTOP data_reprocessing_prompt.	OPEN	RELEASE	03/22/2023 1:53 p.m.	mapr	2881	284
AIRFLOW_data_staging_calibration_sproc4_snapshot20230321	2916	●	Staging tag for data calibrations in sproc4.	OPEN	DEV	03/21/2023 10:25 a.m.	ddossett	3047	62
data_reprocessing_s-proc4_baseline_snapshot20230321	2915	●	Baseline globaltag for the s-proc4 campaign to recalibrate bucket36 with prerelease-07-00-00c. Includes in order of priority patch_main_release-07_noTOP data_reprocessing_prompt.	OPEN	RELEASE	03/21/2023 10:23 a.m.	mapr	2881	284
kfm_time_calibration_cosmics_proc13_prompt	2914	●	Globaltag to be prepended to data_reprocessing_proc13 or data_reprocessing_prompt for reconstructing cosmic runs with release-06	TESTING	DEV	03/23/2023 10:56 a.m.	depietro	24	2
cdcdedx_test_gt_sproc4	2910	●	test gt for sproc4 to validate the constants	TESTING	RELEASE	03/02/2023 4:53 p.m.	renu	279	5
alignment_s-proc4_bucket36ECL	2869	●	Alignment from bucket36 + ECL payloads fix	OPEN	DEV	01/27/2023 11:30 p.m.	bilkat	187	7
bunch_structure_sproc4_v2	2797	●	GT with bunch structure payload for exp26 sproc4	OPEN	DEV	12/07/2022 4:58 p.m.	gpllna	4	1
bunch_structure_sproc4	2796	●	GT with bunch structure payload for exp26 sproc4	OPEN	DEV	12/07/2022 3:06 p.m.	gpllna	7	1
AIRFLOW_final_staging_calibration_sproc4	2795	●	Final staging tag to collect and test all calibrations in sproc4.	TESTING	DEV	03/22/2023 12:46 p.m.	ddossett	3049	66
AIRFLOW_data_staging_localcalib_sproc4	2794	●	Staging tag for local calibrations in sproc4.	TESTING	DEV	12/15/2022 10:26 p.m.	ddossett	14	11

MC Productions

*Lots of run-dependent MC15 skims are available NOW
(the rest will be ready very soon!)

- **Run-independent, e.g. MC15ri_X (moving away from this)**
 - Easier to produce but...
 - → **Beam backgrounds from simulation**
 - Produced in predetermined luminosity (e.g. 1 ab^{-1})
 - Less accurate detector performance and beam backgrounds 🙄
- **Run-dependent, e.g. MC15rd_X (start using this!*)**
 - More difficult to produce (reliant on conditions payloads) but...
 - → **Beam backgrounds from random triggers**
 - Produced in **streams** (1 stream = luminosity of corresponding data)
 - More accurate detector performance and beam backgrounds 👍

Note: Events in MC are NOT rejected according to the L1 or HLT flags

4 streams produced for MC15rd (for BB and qqbar)

- > Belle II Public
- > Belle II Internal
 - > Archive WebHome
 - > Computing Steering Group
 - > Computing WebHome
 - > Data Production WebHome
 - Data production status
 - > Data main page
 - > Offline Luminosity Page
 - > **MC main page**
 - MC run-dependent details
 - MC run-dependent: LowMultiplic
 - MC run dependent signal produc
 - MC run-independent details
 - MC run independent signal prod
 - Instructions to request MC14 Ru
 - > MC production expert page
 - OLD OUTDATED MC15 run index
 - MCri signal misaligned sample
 - > Skim main page
 - Data Production Calibration main p
 - Data Production Validation Page
 - > Data Production Analysis Validatio
 - Data Production service Task list

Pages / ... / Data Production WebHome

Edit Save for later Watching Share ...

MC main page

Good resources!

Umberto Tamponi posted on 11. Mar. 2021 13:29h - last edited by Giovanni Gaudino on 19. May. 2023 16:41h

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
 - 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 result that

Default MC (previously Generic MC)

Default MC is what gets automatically produced in MC production campaigns, and they are just the typical processes which we expect to see at Belle II, such as:

$e^+e^- \rightarrow Y(4S) \rightarrow B^+B^-$ (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)

$e^+e^- \rightarrow llXX$ (eepipi, eepp, etc.) , hhISR (pipISR, KKISR, etc.)

Generated based on central decay file* (one dec file to rule them all...): **DECAY_BELLE2.DEC**



For MCrd we produce 14 types of generic MC:

charged, mixed, $u\bar{u}$, $d\bar{d}$, $c\bar{c}$, $s\bar{s}$,

taupair, ee, mumu, eemumu, eeee, gg, llXX, hhISR

* https://gitlab.desy.de/belle2/software/basf2/-/blob/main/decfiles/dec/DECAY_BELLE2.DEC



Signal MC

Signal MC is specific to your own analysis.

- You can specify it as needed.
- You may need one sample, or multiple different samples.
- Define the decays, branching fractions, decay models, etc.

Dec files: Need to specify your own dec file, named according to the dec file naming rules*

Contact the **DP production liaison** in your working group to get started!

Note: you can also have your signal MC skimmed by your WG **skim liaison**!

* <https://confluence.desy.de/display/BI/Physics+EventType>



Analysis Skims

(dedicated skim talk tomorrow @ 10 am)



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

- Produced as udst (i.e. with particleLists, vertex fit results, etc)
- Preprocessed to save time, small file sizes to save even more time!
- Skims should retain 10% or less of mdst events
- Currently ~70 skims available
- Fully available for data and MC15ri
- MC15rd partially available (almost done!)
- Each WG has a skim liaison

Takeaway: use skims!



See tomorrow's slides for more details!





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... but don't be shy!)

Easy submission to gbasf2

- it is actually **faster** to use Collections compared to using a text file with a list of LPNs

Do **NOT** use the gbasf2 function “-n” to process more than 1 file per job

- Collections contain different campaigns (different globalTags), which cannot be processed together

```
gbasf2 steering.py -p myProjName -i /belle/collection/Data/proc13_had_4S_v3 -s light-2305-korat
```

<https://confluence.desy.de/display/BI/Collection+summary>

Interested in helping?

- Gain experience!
- Authorship qualification!
- Helps the collaboration!
- Good for CV!
- Be more visible in the collaboration!
- Fun!

Lots of ways to help:

1. Leadership role (e.g. skim manager/deputy)
 - a. <https://confluence.desy.de/display/BI/Data+Production+Leadership>
2. DP service task (e.g. a specific project)
 - a. <https://confluence.desy.de/display/BI/Data+Production+service+Task+list>
3. DP production shift
 - a. <https://shift.belle2.org>
 - b. Please try to take these intermittently!



Interested?

Contact Umberto or Stefano!

More questions? Great resources:

Confluence pages: <https://confluence.desy.de/display/BI/Data+Production+WebHome>

B2questions: <https://questions.belle2.org/questions/>

Mailing list: dataprod@belle2.org

Previous Belle II Summer Workshops : <https://indico.belle2.org/event/8841/> (checkout previous DP talks)

Basf2 documentation (Sphinx): <https://software.belle2.org/> (checkout the beginners' tutorial)

Conditions Database: <https://cdbweb.sdcc.bnl.gov/> (globaltag information)

Experiment Numbering: <https://confluence.desy.de/display/BI/Experiment+numbering>

Gitlab (source code): <https://gitlab.desy.de/belle2>

