Summary of PPD contribution from China

Speaker:Muhammad Ahmad on behalf of PPD group

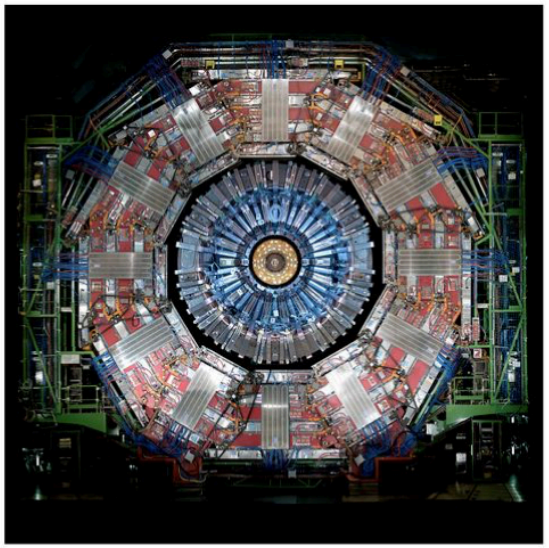
Tsinghua University, Beijing, China

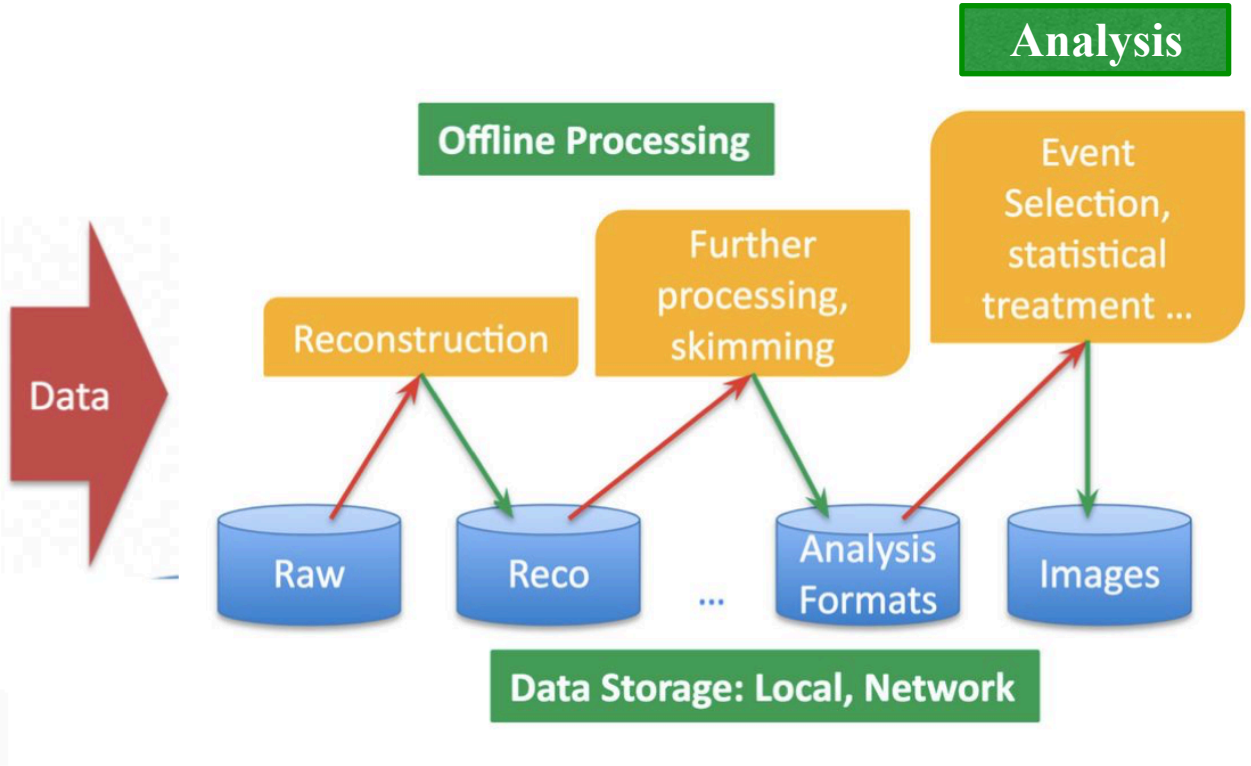
CLHCP 2021 (Nov. 25-28)

Outline

- Aim is to give an overview of extensive work done by Chinese institutes/ universities in CMS PPD coordination area
- PPD organization and mandate
- PdmV organization and mandate
- DQM-DC organization and mandate
  - ✓ New DQM GUI
  - ✓ ML4DQM
  - ✓ Operations 2021 and beyond
  - ✓ Data certification
- Summary

PPD Mandate (link)

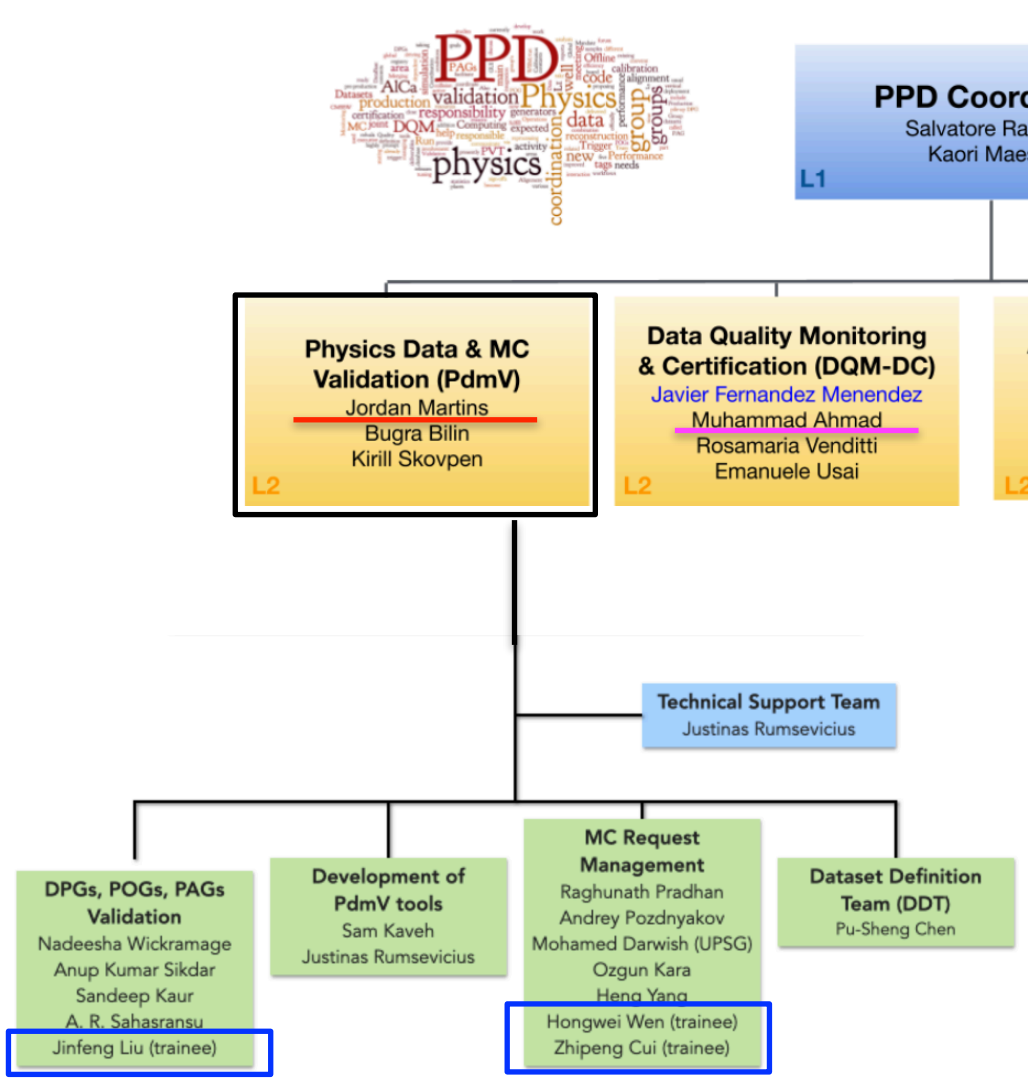




PPD (Physics performance and dataset): Work with other coordination area (Run Coordination, Offline&Comp, Physics) and DPG/POG/PAG…..

- Provide Centralized Online Data monitoring and Data Certification
- Provide Centralized Calibration and Alignment
- Provide Data and MC reconstructed Data for validation and physics in coordination with Physics and O&C
- Validation and Coordination of the CMS reconstruction software

PPD/PdmV group organization (link)



THU/NNU are traditionally leading/involved in PdmV group

L2: Jordan Martins (**2020-2022**)

L3: Jinfeng Liu (**2020-2022**)

L3: Hongwei Wen (**2020-2022**)

L3: Zhipeng Cui (**2020-2022**)

Ex conveners:

L2: Zhen Hu (2018-2019)

L3: Muhammad Ahmad (2019-2020)

PdmV Mendate

- **Data ReReco:** data reprocessing for multiple purposes, with updated conditions
- **Dataset Definition Team (DDT):** definition of the Primary Datasets, including those which are meant for delay reprocessing or data parking. Also, definition and implementation of Secondary Datasets and Skims
- **Development:** improve PdmV management towards its areas by developing and updating relevant tools
- **Release Validation:** samples dedicated to multiple purposes as for the campaigns validation of new CMSSW releases, their fixes and new implementations.
- **Monte Carlo production:** setup of campaigns for central Monte Carlo production and production of datasets according to requests from POGs, PAGs, DPGs.

Main development Projects

- **Production Monitoring:** Stats² and pMpSE
- Dataset Table
- Multi Validation for the root requests in McM
  - All root campaigns are set within the pair up-front (1core,1900MB). McM manages on its own to evaluate the best pair with respect to the CPU eff
- **RelVal machinery**
  - Inspired on McM tool, this is a machinery to replace manual runTheMatrix by manager to deploy relval workflows. Also, centralizes and offer a better bookkeeping
- **ReReco machinery**
  - The same idea of RelVal machinery, but for DATA submissions
- **GrASP**
  - An easy tool to quickly check samples: by physics processes, interested PWGs and tag
  - Ideally, it will be a gDoc replacement toward MC samples preparation in future MC campaigns
- **GPU workflow setup**
- **ExternalGeneratorFilter** implementation (> 10_6_X releases)
  - Designed to take better advantage of CMSSW framework while processing generators not Multi-Thread safe. Ultimately, makes usage of the streams and reduces the time/event. Great application for low filter eff requests

PPD/DQM-DC group organization (link)

PPD Coordination
Salvatore Rappoccio
Kaori Maeshima
L1

Resource Manager
Kirsti Aspola

Database Officer
Giacomo Govi

PC

Physics Data & MC Validation (PdmV)
Jordan Martins
Bugra Bilin
Kirill Skovpen
L2

Data Quality Monitoring & Certification (DQM-DC)
Javier Fernandez Menendez
Muhammad Ahmad
Rosamaria Venditti
Emanuele Usai
L2

Alignment Calibration & Database (AlCaDB)
Francesco Brivio, Helena Malbouisson,
Tamas Vami
L2

Particle Flow (PF)
Kenneth Long
Juska Pekkanen
L2

DQM-DC Conveners
Rosamaria Venditti (INFN-Bari)
Muhammad Ahmad (Tsinghua)
Javier Fernandez (Oviedo)
Emanuele Usai (Brown)

Data Certification

L3 coordinator
Geetanjali Chaudhary (Panjab)
Aliya Nigamova (DESY)

Data Certification
Aayushi Singla (Panjab)

DQM Online Shift Operation
Muhammad Ahmad (Tsinghua)

Online Software Operation
(including ML4DQM integration)

L3 coordinator
Petr Mandrik (Kurchatov)

Development & Support

Software Coordinator
Javier Fernandez (Oviedo)

Online/Offline Developer
Muhammad Ahmad (Tsinghua)
Petr Mandrik (Kurchatov)

Integration
Muhammad Ahmad (Tsinghua)
Petr Mandrik (Kurchatov)

ML4DC
Luka Lambrecht (Ghent)

Tools

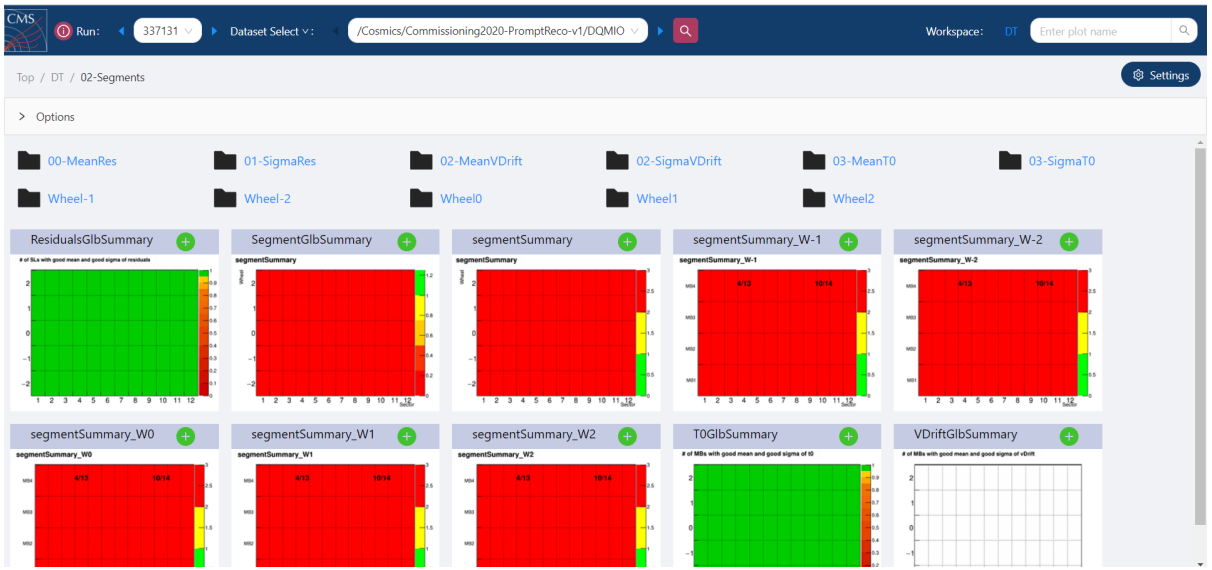
DQM GUI
Chanchana (Chulalongkorn)

Run Registry
Petr Mandrik (Kurchatov)

Historic DQM
Petr Mandrik (Kurchatov)

New DQM GUI

- New DQM GUI has been developed
  - Offline DQM GUI started development in Feb. and it already has beta version (still based on the **old back-end**)
    - https://dqm-gui.web.cern.ch/
  - Online version has been successfully tested during MWGR#3 for the first time (based on **new back-end)**
    - https://cmsweb.cern.ch/dqm/online-new/
  - Commissioning of new (beta version) GUI (running along with old GUI)
  - Received all the runs and few feedbacks from shifters
  - Continuously improving and fixing issues



DQM-DC (ML4DQM, ML4DC)

- Currently data certification is done by humans (for each run)
  - ✳ Prone to errors, person power intensive
  - ✳ Around 50-70 people involved in full process
  - ✳ It is desirable, if Data Quality can be checked LS by LS, rather than run by run
    - ✓ Single LS is data taking in ~23 seconds
    - ✓ In 2017, 2018, each run contains an average of 500 LS and there are around ~200 runs per year
- Machine learning for Offline data certification
  - ✳ Aiming to have a general software toolkit prototype ready by 2nd half of 2021

Run Number	btag	castor	cms	csc	ctpps	lowlumi	dt	ecal	es	egamma	hcal	hlt
297046	GOOD	EXCLUDED	GOOD	GOOD	BAD	BAD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD
297047	GOOD	EXCLUDED	GOOD	GOOD	BAD	BAD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD
297048	GOOD	EXCLUDED	BAD	GOOD	BAD	BAD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD
297049	GOOD	EXCLUDED	BAD	GOOD	BAD	BAD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD
297050	GOOD	EXCLUDED	GOOD	GOOD	BAD	BAD	GOOD	GOOD	GOOD	GOOD	GOOD	B G
297056	GOOD	EXCLUDED	GOOD	GOOD	BAD	BAD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD
297057	GOOD	EXCLUDED	GOOD	GOOD	BAD	BAD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD

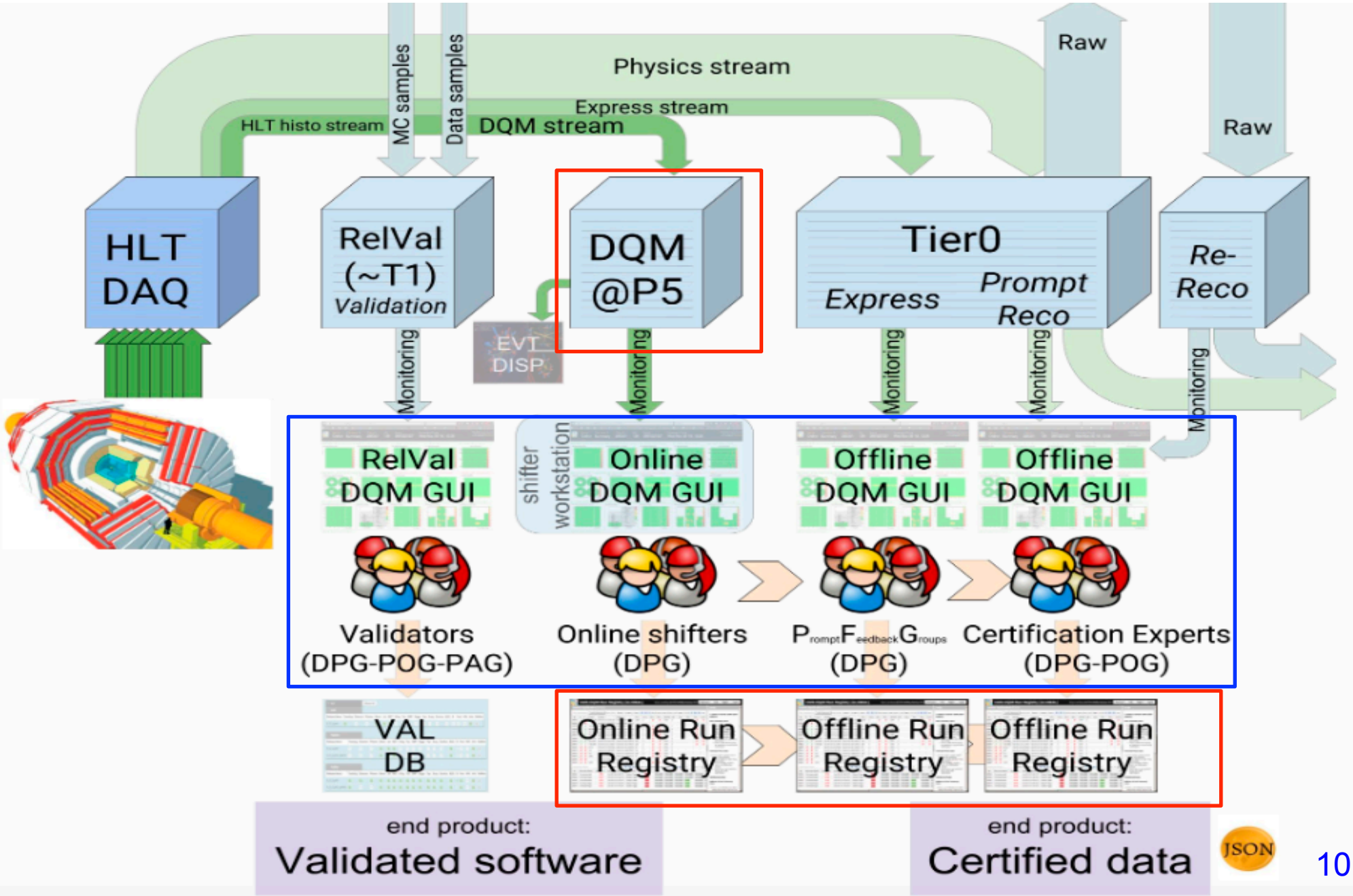
- Software toolkit to support the subsystems in ML4DC
  - Utilities for definition of dataset, ML model choice, training/test, resampling tool to generate bad examples
  - Combination of outputs on several histograms in a single flag
  - Integration in DQM GUI and final flag in the Run Registry
- Project just started (The toolkit is still in a design stage)
  - A prototype working on a single subsystem (pixel) exists
  - Code and documentation in development

Technical points to be addressed:

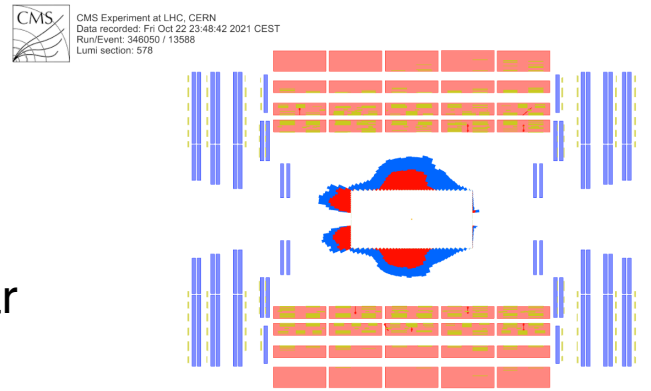
-Storage of per-LS files: need **nanoDQMIO**?

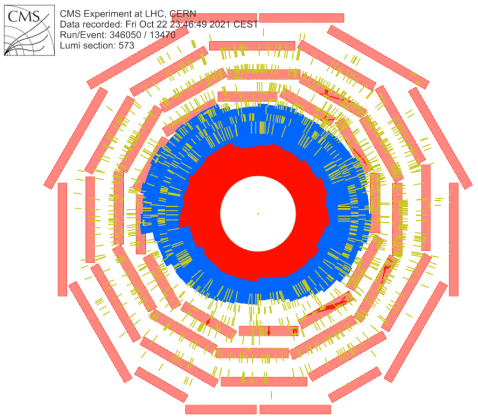
- Produce DQMIO just on the **4 primary dataset** used for data certification

DQM system at point point 5



Operations 2021 - shift organization





- ONLINE DQM shifts in remote mode since 2020, keep the remote mode at least until April 2022
- Arranged many tutorials for shifter around the year
- Put significant efforts to adopt smooth DQM operations from remote
  - 4 MWGRs in 2021
  - July/Aug: 5 week of CRUZET, 24/7 operations
  - 1 week CRAFT (oct), 2 week of test beam (splashes, collision)

- DQM Code optimization to evaluate CPU/timing performance:
  - Main sources of consumption found and checked by subsystems and significant improvement has been done. Also DQM Matrix (running on T0 for prompt reco.) have optimised for PDs and subsystem
- Central Data Certification (DC)
  - The DC process has been established again after Run2
  - Performed DC for CRUZET, report can be found here [Link]
  - Performed DC for CRAFT and test beam, report can be found here [Link]

Summary

- Extensive work done (& in progress) on various fronts for coming Run3 preparations in PPD coordination area by THU/NNU
- **Contributing in PdmV:** Responsible for CMSSW validation, MC production, development of relevant tools and data reprocessing for CMS collaboration
- **Contributing in DQM-DC: R**esponsible for CMSSW DQM code, DQM operations and quality maintenance of data , development of relevant tools (DQM GUI, RR etc) and Data certification
- These contributions with leading roles in groups: Significant visibility of Chinese community in CMS Collaboration