Data Quality and Certification Plans for 2017

CMS Run & DPG Coordination Workshop
January 26, 2017

Virginia Azzolini (Massachusetts Inst. of Technology, US)
Ridhi Chawla (Panjab University (IN))
Anterpreet Kaur (Panjab University (IN))

On behalf of Data Certification - DQM team

Outline

What we did: brief summary about 2016 data certification

Identified 3 main areas of interest

- Tools
- Human element
- Workflow

For each of them

- What we learnt: areas of improvements
- What's next: preparation for 2017 operations

Summary

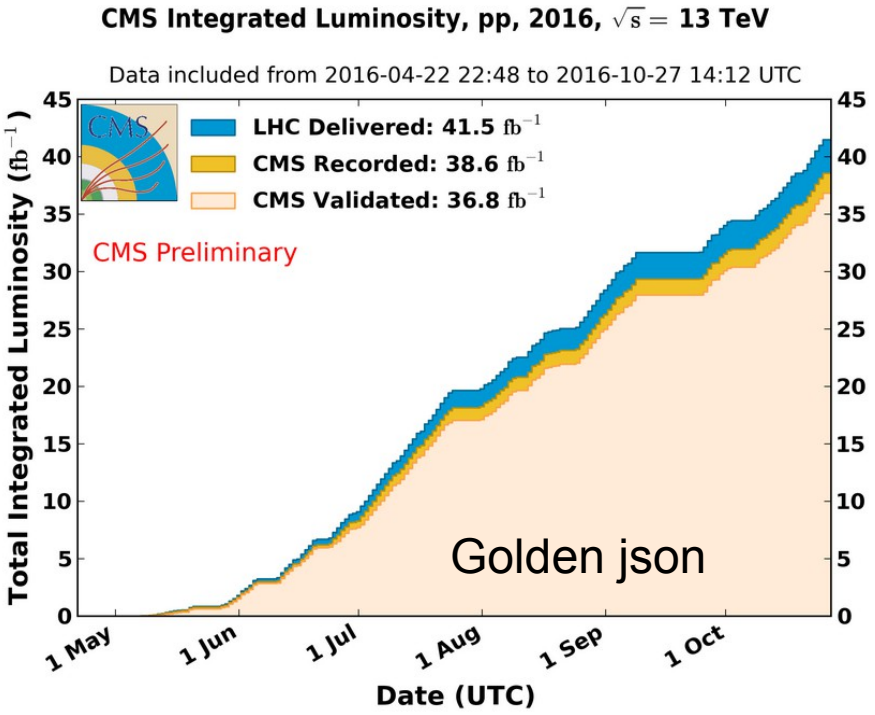
2016 pp Collisions in numbers

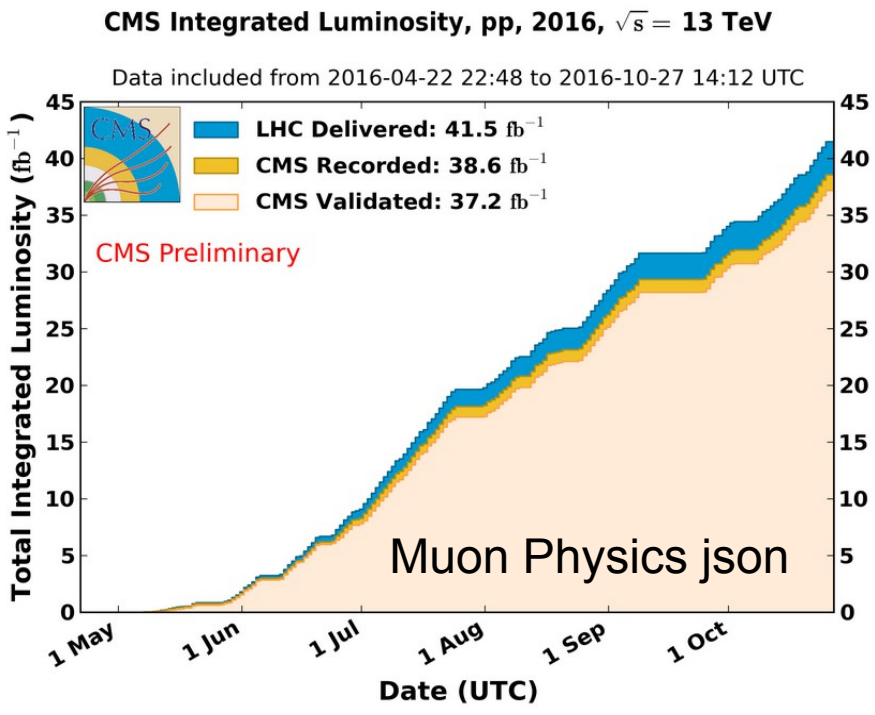
Status at Jan 10th, 2017

. Standard physics **eraBCDEFG 23 Sep Rereco** + **eraH Prompt Reco**

95.6 % data taking efficiency

96.5% Good data efficiency





. **Low PU physics**, set of runs with homogeneous **PU~1** : 1.63 /pb

Note: in the above plot, only runs potentially good for pp Physics are considered.
Runs for VdM scan, Mu scan, beam background studies, with non standard magnetic field value are not considered. in the Short runs (< 80/nb) are not considered for DC since DPG/POG needs enough stat for providing reliable feedback

2016 pPbp Collisions in numbers

Status at Dec 14th, 2016

Era	run min	run max	E and beams	Golden Lumi	Teff	Geff	Muon Lumi	Teff	Geff
PARun2016B	284756	285396	5 TeV p-Pb	145.87 /ub	97%	36%	176.68 /ub	97%	44%
PARun2016C	285410	285951	8 TeV p-Pb	64.41 /nb	97%	99%	64.41 /nb	97%	99%
PARun2016C	285952	286496	8 TeV Pb-p	115.28 /nb	96%	99%	115.2 /nb	96%	99%
PARun2016D	286506	286520	5 TeV Pb-p	104.06 /ub	94%	100%	104.06/ub	94%	100%

Same runs with Castor feedback included

Era	run min	run max	E and beams	Golden Lumi	Teff	Geff	
PARun2016B	284756	285396	5 TeV p-Pb	145.87 /ub	97%	36 %	same as golden standard json
PARun2016C	285410	285951	8 TeV p-Pb	64.38 /nb	97%	98.9%	
PARun2016C	285952	286496	8 TeV Pb-p	115.28 /nb	96%	99%	same as golden standard json
PARun2016D	286506	286520	5 TeV Pb-p	96.02 /ub	94%	93%	

Teff: data taking efficiency
Geff: Good data efficiency

Tools

. Data Certification

We learnt

.. scripts: easy configurable, but require manual intervention → error prone

.. full control requires many little steps, perhaps too many

For the next

.. aiming for DC **tools** easier to use, **less needy of previous understanding**

For the next next tools **more intelligent** (supervised or unsupervised NN)

. Run Registry and WBM: story of a difficult relationship

.. RR, adopted by WBM, is de facto still a creature in the hand of few → urgency happen

.. become a better consumer of **jira tickets** for solve technical and operational issues

. Run Registry: One Tool to rule them all

.. impose to add a **reason** request, when a run is marked **BAD** (now kindly asked, sometimes is forgot)

.. ask DPG/POG experts to **justify** to DC **changes applied a posteriori** to a run (both GOOD and BAD)
Once a run is in a JSON file, any change will impact the work of the analists, they deserve a reason

.. ask groups to start to use or make a better use of the **RR feature cause***
This would allow
. pre-thinking : find common peculiarities in data for standardizing DC outcome
. routinely: tag them automatically, possibility to concentrate DC only on unusual cases
. a posteriori: collect statistics about problems of same type via the imposed wild card

* https://twiki.cern.ch/twiki/bin/viewauth/CMS/DQMRunRegistry
https://twiki.cern.ch/twiki/bin/view/CMS/TutorialRRforDQM

Human element

. Communications

.. many private conversation instead of HN or egroups → people/discussion is cut off & loss history

. No Communications

.. e-groups, preferably used by DC, often do not reply to our emails

. Lack of certifiers/reports:

.. Electrons (Egamma group) had no certifiers for some time → checks missed (found one for 2017)

.. L1T lack of man power → substantial delay in DC → extra work for all

For the next

.. **identify, meet and know better the DC responsibles**

. Review together their workflow and tools (some are still using private code running)

. Identify the responsible for a final word, in case needed. Possible in CERN timezone.

.. re-organization of DC **experts in shifters (Doc3)**

. round-the-clock certification

. depersonalize the DC outcome, making it general and absolute

. possibility to gain central shift EPR point for more people

. need to identify a constant presence(shift leader), reference & replier to posteriori or difficult questions

Workflow

. Workflow

.. easy to run in routine time, hardly scale to unusual time → gather info can be a nightmare
Info too scattered in several sites
and Elogs often cryptic or incomplete

.. update DC documentation, sharing and adding lessons learnt and linking source of interesting info
.. emphasize the need of have as many info possible in Elogs for future reference

. report deadlines often missed

.. ideally: experts look at runs → DC produce json → no changes needed after presentation talks
.. ideally: DPG deadline should be internally set few h in advance
→ allow the POG to, after have done their validation, check the dependencies wrt to DPG
In case of problem, have enough time to think about it, contact the DPG related
and still be in time to make the midnight thursday deadline.
.. automatization: once that RR will be able to freeze reports status at midnight,
all these delays will render this possibility useless and prevent some automation of the process

.. wondering:
. just a bad habit or is there a real kink in the process that needs to be smoothed?
. should we treat these delays as a show stopper? how much strict could and should be?
. talk to DC certifier/responsible will do ?
. change the workflow will benefit the task? (see next)

Workflow

.. change the workflow of weekly certification – proposal*

2016:

S	M	T	W	T	F	S	
	.. take data	.. take data	.. take data	.. take data	.. take data	.. take data	Week 1
.. take data	Circulate run list W1		Freeze run list W1	Reports midnight deadline W1	Certification verification and json prod W1		Week 2
		Report run coord json W1		Report PPD W1			Week 3

2017:

S	M	T	W	T	F	S	
				.. take data	.. take data	.. take data	Week 0
.. take data	.. take data	.. take data	.. take data	Circulate run list W 0-1			Week 1
	freeze run list W 0-1 ---------- Reports midnight deadline W 0-1	Certification verification and json prod W 0-1		Report PPD W 0-1			Week 2
		Report run coord json W0-1					Week 3

* first attempt

Workflow

S	M	T	W	T	F	S	
				.. take data	.. take data	.. take data	Week 0
.. take data	.. take data	.. take data	.. take data	Circulate run list W 0-1			Week 1
	freeze run list W 0-1 ---------- Reports midnight deadline W 0-1	Certification verification and json prod W 0-1		Report PPD W 0-1			Week 2
		Report run coord json W0-1					Week 3

. change the workflow of certification

.. reports:

	Json outcome	data taken
PPD	1 day old (5)	up to 7 d before(10d old data)
Run coord:	6 day old (3)	up to 12 d (8d old data)

Pro: report to PPD will be more fresh

Con: report to run coord less efficient to prevent damages to data .

Prompt feedback group, and their daily certification of sampled express stream data*, could help.

Pro: tighter relation between PFG and offline DC

.. change of the function of “freeze run list email” :

Now: used as a starting moment of activities (learnt monitoring during 2016)

2017: finalize or verify that all the work for the week is done

Easy to with a continuos certification or attention → Pro: reduce delay

.. slim of operation:

No more circulate run list warning. Usefulness not proved, no related activity noticed

Pro Pro: foreseen an empty day (wednesday) for possible extra checks. No more _noXYsubsystem Json

https://cms-service-dqm.web.cern.ch/cms-service-dqm/CAF/certification/Collisions16/13TeV/CertSummary/status.Collisions16.html

Summary

We consider 2016 a successful year in term of data taking and data certification

Could 2017 be the same?
    of course we could do the same and still succeed

Should we act in 2017 as in 2016?
    Perhaps not

Identified few fields of improvements
    . tools
    . humans
    . workflow

2017 just started, we have few months to work in those areas...
    any comment will be welcome

Please address your email to e-group: cms-dqm-data-certification
Thank you

Backup

S	M	T	W	T	F	S	
	.. take data	.. take data	.. take data	.. take data	.. take data	.. take data	Week 1
.. take data	Circulate run list W1		Freeze run list W1	Reports midnight deadline W1	Certification verification and json prod W1		Week 2
		Report run coord json W1		Report PPD W1			Week 3

S	M	T	W	T	F	S	
				.. take data	.. take data	.. take data	Week 0
.. take data	.. take data	.. take data	.. take data	Circulate run list W 0-1			Week 1
	freeze run list W 0-1 ---------- Reports midnight deadline W 0-1	Certification verification and json prod W 0-1		Report PPD W 0-1			Week 2
		Report run coord json W0-1					Week 3