CMS COMPUTING & SOFTWARE FOR ANALYSIS

A GENERAL OVERVIEW

Tommaso Boccali

INFN Pisa

CMS Computing model

- The Computing Model (and the tool designed to implement it) is the way we have designed to allow **USERS** to do **PHYSICS** at CMS.
- It covers every aspect of the processing and analysis, from the moment in which Data was saved by CMS @ P5 to the approval of an analysis paper.

- It is indeed complicated, by why so complicated?
- Some numbers follow …

CMS offline in numbers

- CMS (RunI) was typically taking data @ 500 Hz, which results in ~ **5 Billion** interesting events per year (already removing calibration streams)
- To do physics with these, we need Simulation (Monte Carlo) samples to compare, typically another ~ **10 Billion Events** / year

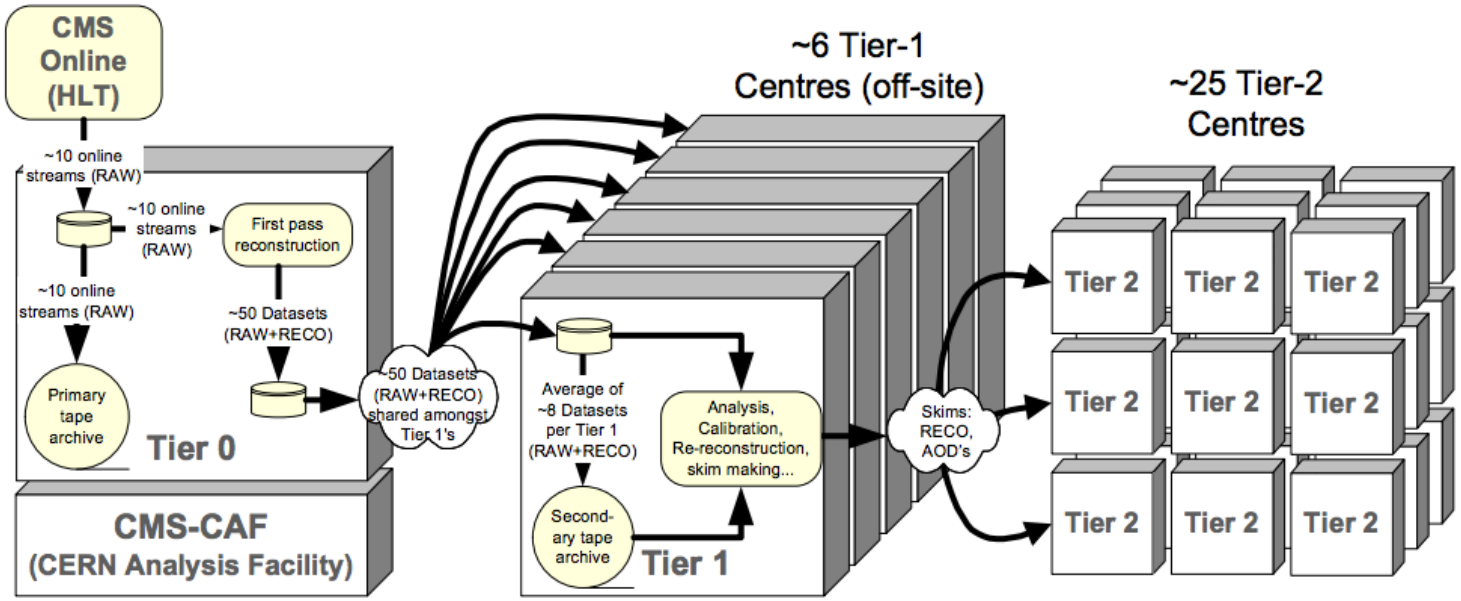
- Both need
  - Computing resource for reconstructing, reprocessing, analyzing
  - Storage space in various centers

- **CMS Today (world aggregate figures)**
  - **~100.000 CPUs of computing power**
  - **~30.000 Hard Drives**
  - **~30.000 Tape Drives**

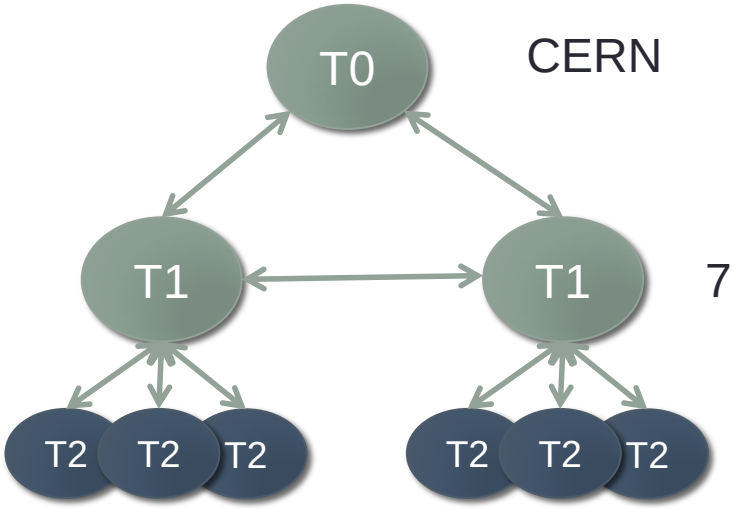
The CMS Computing infrastructure

- Such numbers cannot be managed in a single place (CERN), hence we use a **GRID model** with (main) ~ **60 centers worldwide**; the model is hierarchical
- **1 Tier 0**: CERN, where data is taken and processed "promptly"
  - CERN produces the "Prompt" version of all the samples within 48 hours
- **~7 Tier 1s**: they handle reprocessing and MC production
- **~45 Tier 2s: they handle MC production and analysis tasks**

- We also have local analysis centers, which are not in the model. They can be whatever, from very big (FNAL/LPC, CERN/CAF), to your desktop …



Picture from 2005 Computing TDR



Users should know/use just the T2s. All the rest is used in transparent computing operations

The GRID paradigm

- A “GRID” paradigm is used to link all the resources together. The system should present itself to YOU as a **single entity**, and the sparse location of resources should be invisible
- **This means to you**
  - You **can and should** suppose all the CMS data and resources are within your reach, and you do not need to know where they are
  - The size of your local resources should not matter: every CMS user is on equal ground when doing physics analysis
  - The complexity needed to run the distributing computing effort should be shielded to the user
- (well, at least that was the idea; actual implementation can need a bit of “deeper understanding” on the user side)

Doing analysis at CMS

- **Physics environment is complicated** (20+ interactions over imposed to your signal)
- **Detector is complicated** (~ 100 Million acquisition channels)
- **Computing is complicated** (huge resources, sparsely located)

- Fortunately, there are tools which try and help in analysis activities ... I will describe here the ones helping with the last point

Usual schedule of an analysis @ CMS

- **Run on DATA/MC and produce either**
  1. Simple skims (select events and save only for those a subset of the AOD/MiniAOD content) – very simple
  2. EDM files containing new ad-hoc DataFormats – complicated, needs deep understanding of the CMS EDM FrameWork
  3. ROOT plain Tuples (possibly created using EDM services)

- **Analyze the output**
  - If you chose 1. or 2., most probably the final step of the analysis will take place in FWLite or PyROOT (which is ROOT + CMS object definition either in C++ or Python)
  - If you chose 3., using ROOT/PyROOT as-is will be ok.

- This is typically a choice of the analysis group, it is not clear **YOU** can choose

(a parenthesis)

- ALL CMS jobs usually produce are ROOT files; but not all ROOT files are equal ;)
- If they contain just standard ROOT objects (TH1F, RootTuples) they can be read with standard ROOT
- CMS processed files are instead **E**vent**D**ata**M**odel ROOT files; they contain dumps of C++ objects, and standard ROOT cannot interpret them (or, only partly)
- To correctly use them, you need to “teach” ROOT how to handle these (explain which are the C++ Objects saved); this can happen
  - **If you use CMSSW jobs**
  - **If you use ROOT extension called FWLite**

Analysis at CMS…

- The last part of an analysis can vary a lot (fits, cuts' optimization, …) so cannot be covered in a talk like this
- **But, inevitably, every CMS analysis will have to run on data and most probably Monte Carlo simulation**

- **This is a tremendous task**: running on 1B Events (typical size of a data dataset), even at 0.1 sec/ev, needs 100Msec = 1500 days = 5 years
  - You can only do it using many computers at the same time
- Moreover, those 1B events are probably > 200 TB, so they do not fit a single modern computer (which anyhow would not be able to serve many parallel processing jobs)

- Users definitely need help on this task …

What is needed to perform a processing pass on DT/MC@ CMS?

1. **Know the physics** (this is up to you, guys…)
   - Which means prepare the program/strategy which selects the signal you want to study
2. **Identify the correct data to use**
3. **Identify the correct MC** (if needed, probably yes) **to use**
4. **Find Data and MC** (== where they are located)
5. **Find the correct version of the software**
6. **Find the correct version of the calibration**
7. **RUN on it**
8. (…)

- Let's try to explain the tools at least for 2.-7.
- But before: you starting point for everything are
  - The CMS Workbook (here): general level, usually "from scratch"; contains also info similar to this talk
  - The CMS Software Guide (here): software documentation, use it to write/modify existing software

2. Identify the data

- Data collected by CMS is organized into **Datasets**: sets of uniform files (close triggers, same MC generators, etc)
- A dataset is uniquely identified by text fields, divided by '/'
  - **/ PRIMARYDATASET / ADDITIONALINFO / DATA**
- **Primary Dataset (PD) identifies the lo the dataset**
  - For DATA physics obj trigger. Exam **SingleElectro n, Jets**, etc etc
  - For MC, it contains the basic process information, and the generator used
    - **Ttbar, QCD**, etc etc with Pythia, tauola, etc

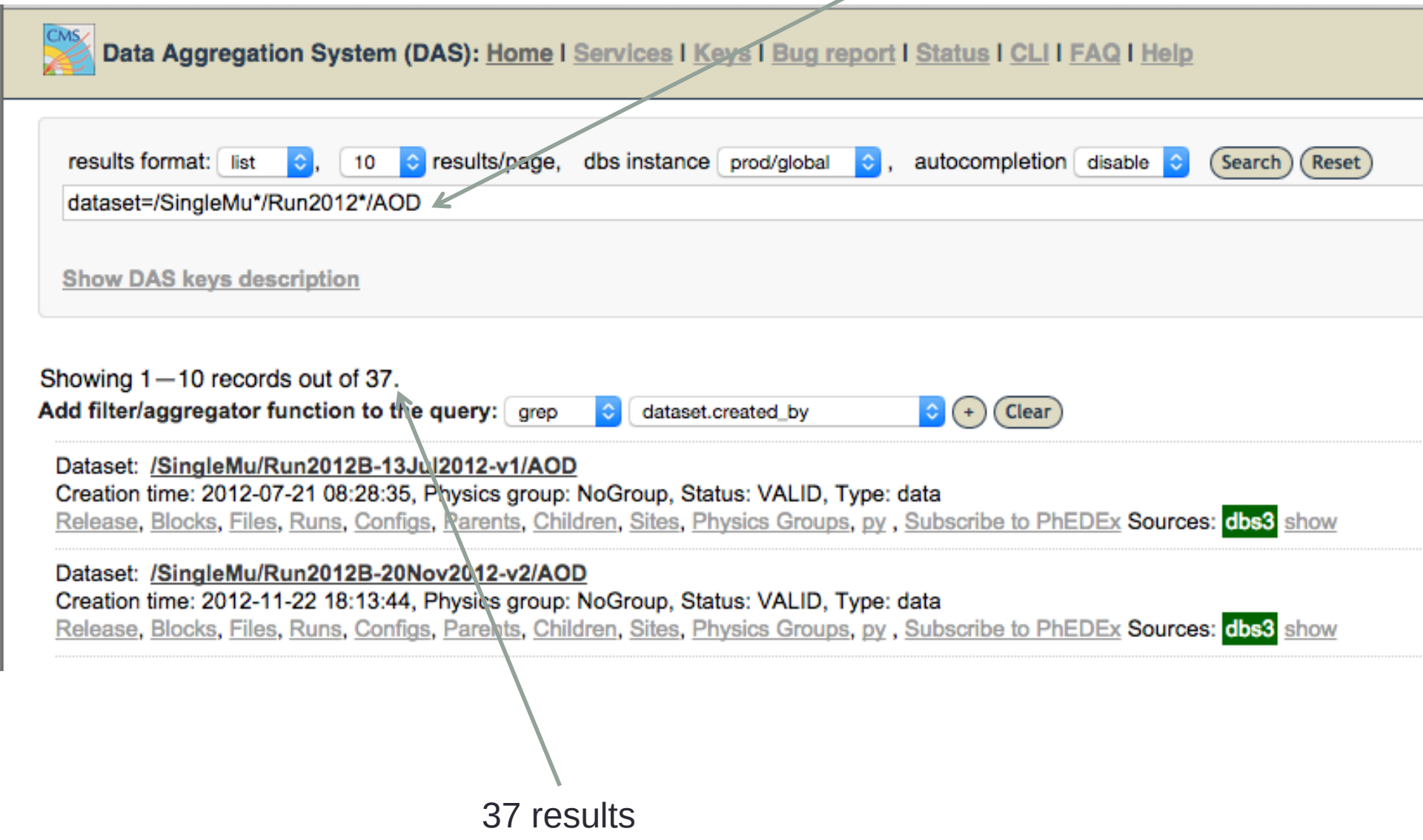
What data is available then?

- **Each PD is present in more data TIERS**
  - **RA** electronics response (useless for
  - tructed events (ok for analysis but ED)
  - usually ok for analysis
  - pact than AOD
  - u add SIM – i.e. AODSIM)
  - o the generator information
- **nfo contains many things**
  - e used for the processing
  - ata / MC campaign (i.e. Summer12)
  - Processing number
  - …
- For example, data taken in 2012 with the SingleMu PD, in AOD format, will contain (at least)
  - PD = SingleMu
  - Period = Run2012
  - DataTier = AOD

Use Data Aggegation System (DAS)

- Here: https://cmsweb.cern.ch/das/
- Search for something like **/SingleMu*/Run2012*/AOD** (with wildcards!)
- You will get 37 answers; the most relevant are
  - Those like **/SingleMu/Run2012C-PromptReco-v1/AOD**
    - The **"C"** is the third data taking period of the year
    - **"PromptReco"** means the sample was produced @ CERN Tier0
    - **"v1"** is a smaller specification wrt C
  - Those like **/SingleMu/Run2012A-23May2012-v2/AOD**
    - This is a reprocessing of the data, which superseded PromptReco samples (it is the reprocessing launched on May 23rd)
- But there are many, which ones are really relevant?

Your query



37 results

Some help …

- https://twiki.cern.ch/twiki/bin/view/CMS/Collisions2011Analysis
- https://twiki.cern.ch/twiki/bin/view/CMS/PdmV2012Analysis
- They contain the **last + exclusive + complete** let of datasets you need to use to access a full period

DATASET	from Run	to Run
Run2012A-22Jan2013	190456	193621
Run2012B-22Jan2013	193833	196531
Run2012C-22Jan2013	198022	203742
Run2012D-22Jan2013	203777	208686

So you will need:

/SingleMu*Run2012*-22Jan2013/AOD

…

Not enough …

- Not all the data in these Datasets is good: sometimes detectors go off, trip, we have problem with DAQ, we lose magnetic field …
- A run can last 12 hours … many things can go bad, even for a few seconds
- So the run cannot be the smallest event unit we have!
- **LS = Luminosity Section**
  - It is the quantum of the data taking, it lasted 23 sec in RunI
  - **We ASSUME nothing changes abruptly during a LS, so all the events inside are ASSUMED to have been taken in identical conditions (detector conditions, accelerator conditions etc)**
- **Certification works at LS level: a LS can be either all good or all bad**

A short PS: assume means

- We do not change anything willingly
- We hope conditions not under our control change too much

JSON selection

- For each DATA dataset, there is a list of GOOD/BAD runs which is maintained by the Physics Performance & Datasets (PPD) Project.
  - They are certified by detector experts
  - The lists come in the for of JSON files


```
{"190645": [[10, 110]],
 "190704": [[1, 3]],
 "190705": [[1, 5], [7, 76], [78, 336], [338, 350], [353, 384]],
 "190738": [[1, 130], [133, 226], [229, 355]],
```


  - Which means: from Run 190645 you can use LSs from 10 to 110, etc etc

  - The official list is either in the table on the twiki, or in
    - **/afs/cern.ch/cms/CAF/CMSCOMM/COMM_DQM/certification/Collisions12** or similar

So all in all, to IDENTIFY data

- You need a list of **comprehensive + exclusive** datasets for the data taking period you are interested in

- You need a list of JSONs file from Certification
  - Or a single one including all the periods

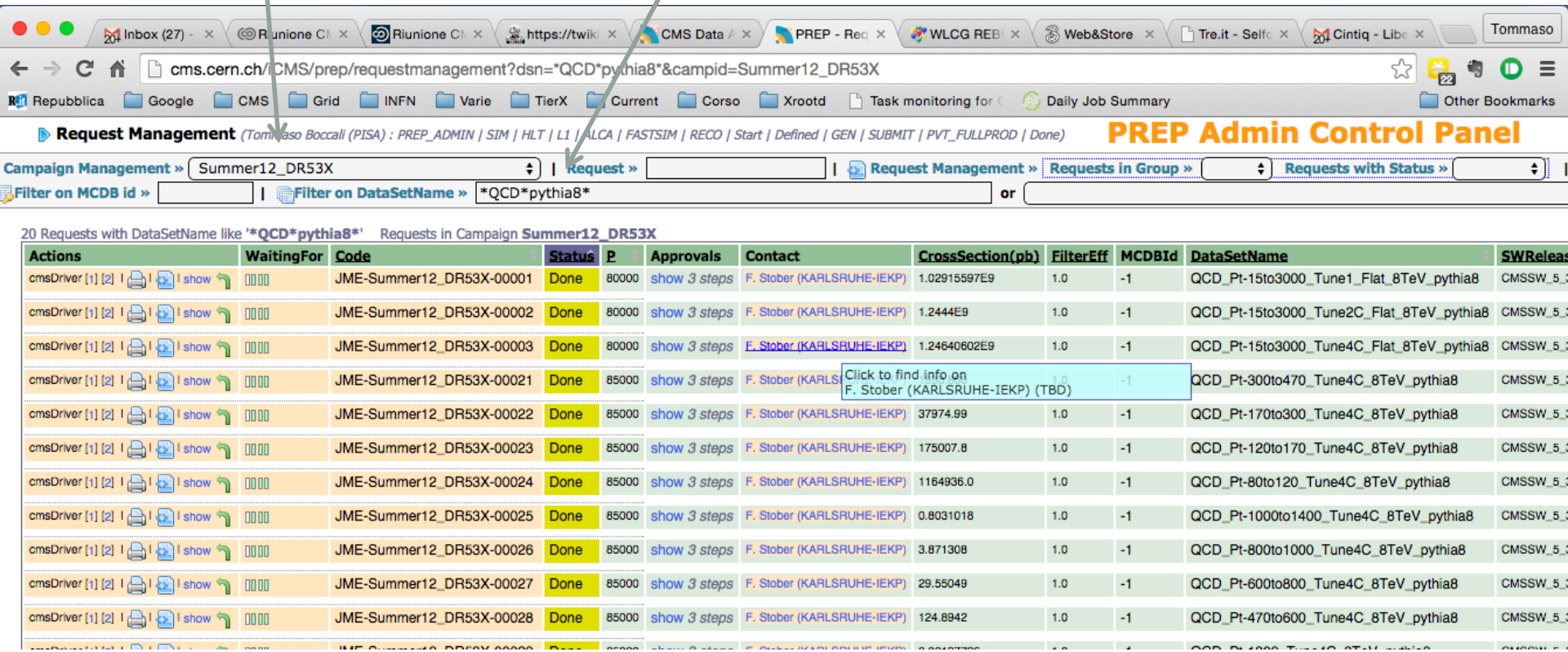
  - Jobs are intended to run on events/LSs which are in the dataset + selected in the JSON

3.Identify MC

- A similar task than with data, but
  - There are no bad events: no need for a JSON file
  - The same sample usually exists with different pileup (lumi) conditions – choose the right one
- Your entry point for RunI is PREP (http://cms.cern.ch/iCMS/prep/requestmanagement)
  - Usually you specify Campaign name + put a query on dataset name
  - For example for 2012 CMSSW53X, QCD samples with Pythia8, you choose
    - **Campaign = Summer12_DR53X**
    - **Datasetname = *QCD***

Campaign

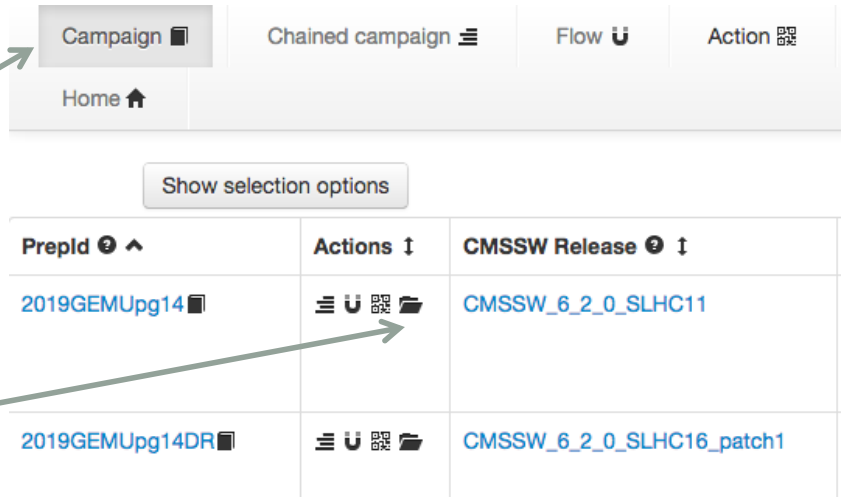
Dataset

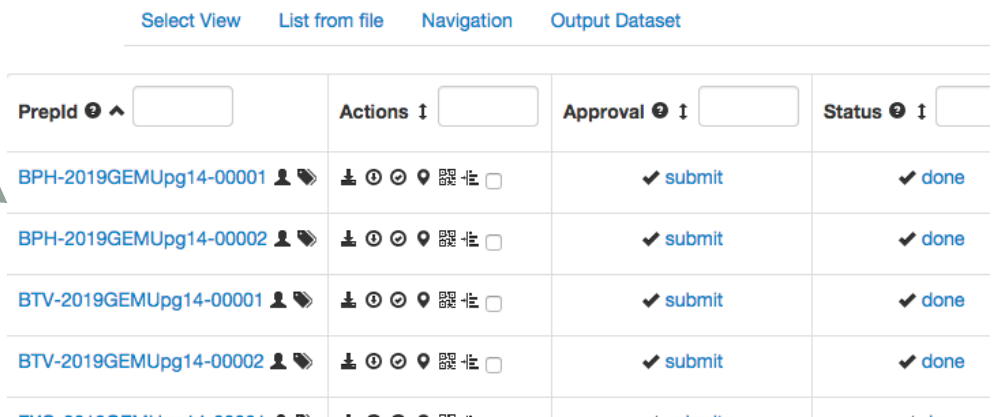


Actions	WaitingFor	Code	Status	P	Approvals	Contact	CrossSection(pb)	FilterEff	MCDBId	DataSetName	SWReleas
cmsDriver [1] [2] I show		JME-Summer12_DR53X-00001	Done	80000	show 3 steps	F. Stober (KARLSRUHE-IEKP)	1.02915597E9	1.0	-1	QCD_Pt-15to3000_Tune1_Flat_8TeV_pythia8	CMSSW_5_3
cmsDriver [1] [2] I show		JME-Summer12_DR53X-00002	Done	80000	show 3 steps	F. Stober (KARLSRUHE-IEKP)	1.2444E9	1.0	-1	QCD_Pt-15to3000_Tune2C_Flat_8TeV_pythia8	CMSSW_5_3
cmsDriver [1] [2] I show		JME-Summer12_DR53X-00003	Done	80000	show 3 steps	F. Stober (KARLSRUHE-IEKP)	1.24640602E9	1.0	-1	QCD_Pt-15to3000_Tune4C_Flat_8TeV_pythia8	CMSSW_5_3
cmsDriver [1] [2] I show		JME-Summer12_DR53X-00021	Done	85000	show 3 steps	F. Stober (KARLS		1.0	-1	QCD_Pt-300to470_Tune4C_8TeV_pythia8	CMSSW_5_3
cmsDriver [1] [2] I show		JME-Summer12_DR53X-00022	Done	85000	show 3 steps	F. Stober (KARLSRUHE-IEKP)	37974.99	1.0	-1	QCD_Pt-170to300_Tune4C_8TeV_pythia8	CMSSW_5_3
cmsDriver [1] [2] I show		JME-Summer12_DR53X-00023	Done	85000	show 3 steps	F. Stober (KARLSRUHE-IEKP)	175007.8	1.0	-1	QCD_Pt-120to170_Tune4C_8TeV_pythia8	CMSSW_5_3
cmsDriver [1] [2] I show		JME-Summer12_DR53X-00024	Done	85000	show 3 steps	F. Stober (KARLSRUHE-IEKP)	1164936.0	1.0	-1	QCD_Pt-80to120_Tune4C_8TeV_pythia8	CMSSW_5_3
cmsDriver [1] [2] I show		JME-Summer12_DR53X-00025	Done	85000	show 3 steps	F. Stober (KARLSRUHE-IEKP)	0.8031018	1.0	-1	QCD_Pt-1000to1400_Tune4C_8TeV_pythia8	CMSSW_5_3
cmsDriver [1] [2] I show		JME-Summer12_DR53X-00026	Done	85000	show 3 steps	F. Stober (KARLSRUHE-IEKP)	3.871308	1.0	-1	QCD_Pt-800to1000_Tune4C_8TeV_pythia8	CMSSW_5_3
cmsDriver [1] [2] I show		JME-Summer12_DR53X-00027	Done	85000	show 3 steps	F. Stober (KARLSRUHE-IEKP)	29.55049	1.0	-1	QCD_Pt-600to800_Tune4C_8TeV_pythia8	CMSSW_5_3
cmsDriver [1] [2] I show		JME-Summer12_DR53X-00028	Done	85000	show 3 steps	F. Stober (KARLSRUHE-IEKP)	124.8942	1.0	-1	QCD_Pt-470to600_Tune4C_8TeV_pythia8	CMSSW_5_3

PREP was used for RunI

- For RunII, and upgrade studies, **McM** is being used (https://cms-pdmv.cern.ch/mcm/)
  - Choose campaign name
  - In a campaign, list the datasets





4. Find data and MC

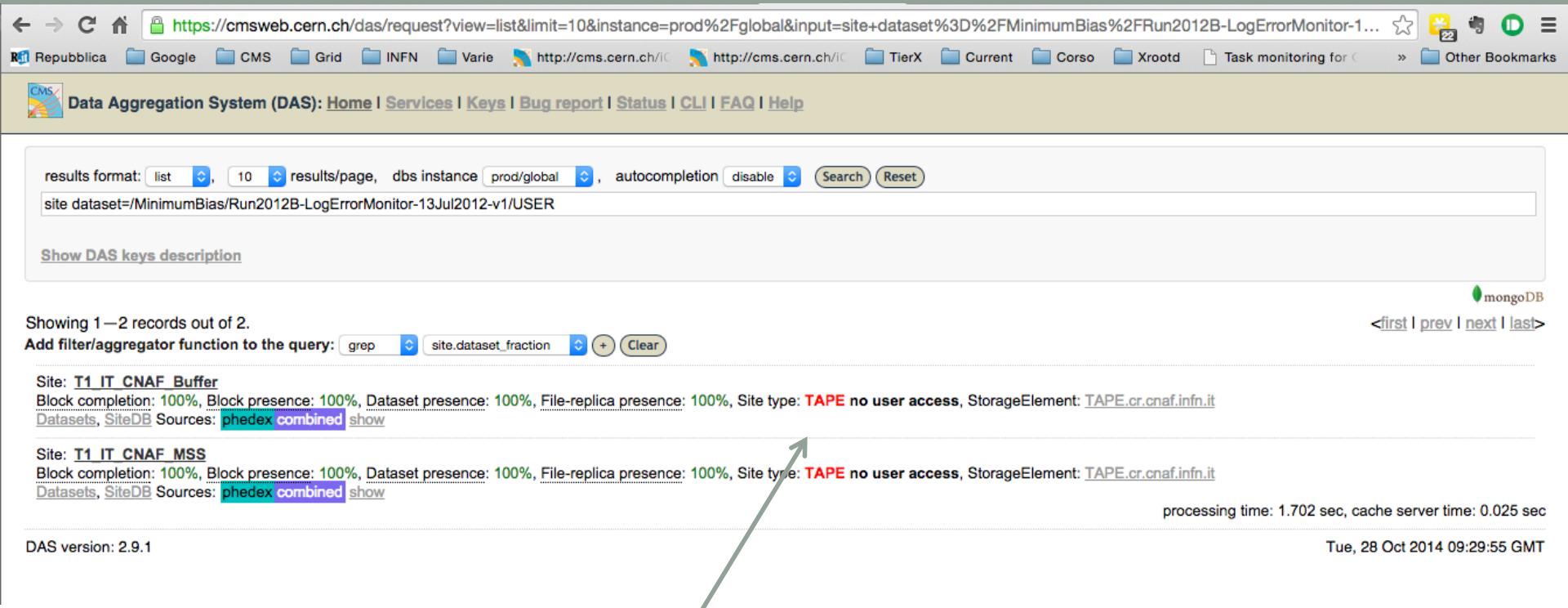
- Ok, let's assume also here you got a list of datasets to process, both for signal and background

- A side note: it is important (unless you really know what you are doing) to get compatible datasets for data and MC
  - In a nutshell: processed with the same reconstruction code

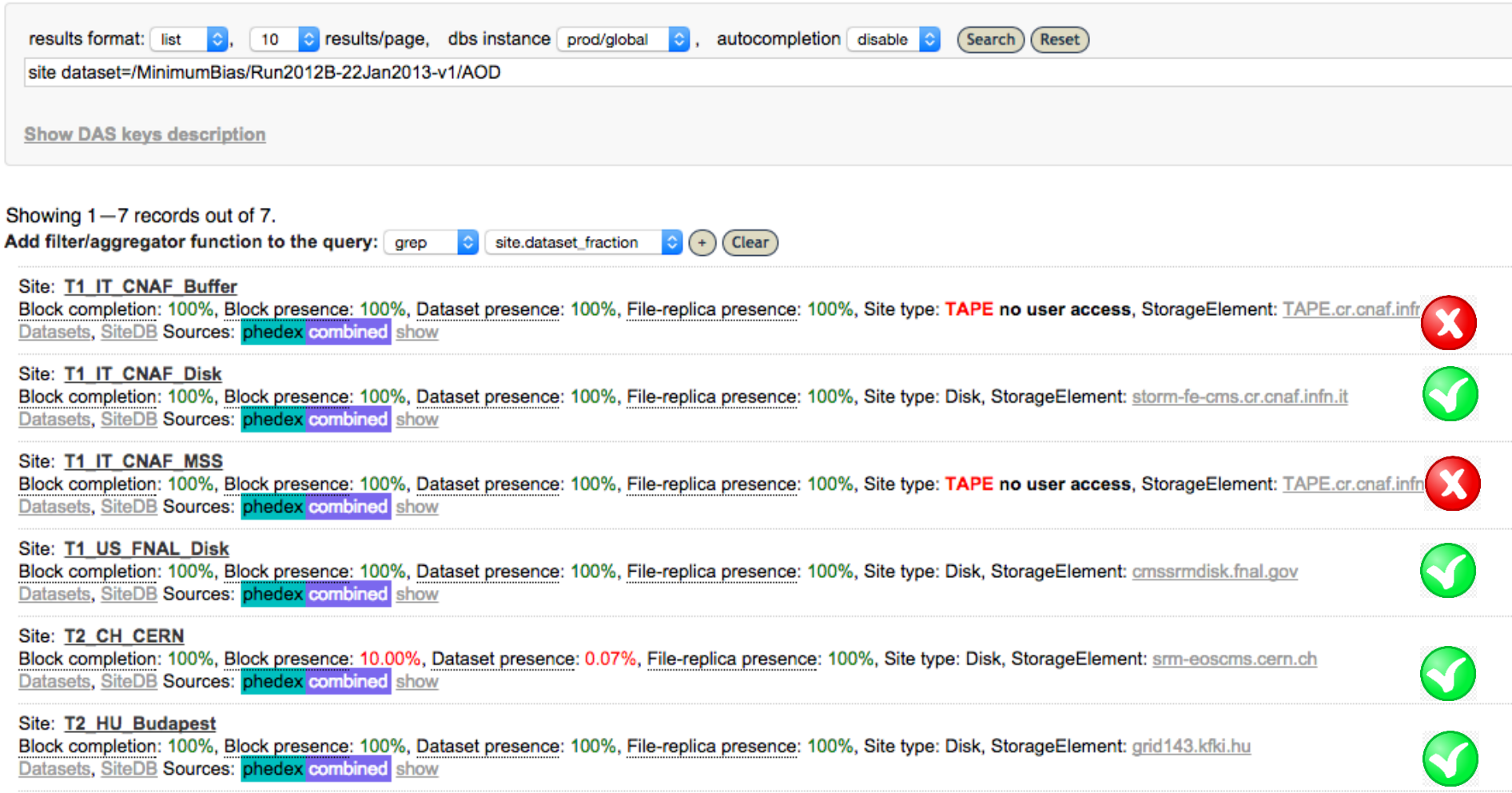
- The twikis linked before should help you in this
- Otherwise, just choose samples processed with as-close-as-possible CMSSW releases (see later)

Find the data …

- At this point, you should understand **where** datasets really sit in the CMS computing hierarchy …
- **Should you? As said, in first approximation you should not care, the system should**
- Nevertheless, if you can make sure a dataset is in more than one place YOU CAN ACCESS, it will make your processing faster and more reliable
- **Where can you access data and run jobs?**
  - At all T2s, identified by T2_XX_YYYYY
  - At Tier1s, but only if the sample is on the Disk area: T1_XX_YYYYY_Disk
- **Use again DAS:**
  - `site dataset=DATASETNAME`
  - Will return you the sites where it is
  - **If in no appropriate place, you can issue a request to move it in one**.



For example, this dataset chosen randomly is NOT ok: it only on TAPE, and DAS tries to warn you ….



So, overall the sample is well accessible…

5.Find the correct version of the SW

- CMS uses for “almost everything” CMSSW as the application software
  - From trigger, to reconstruction, to analysis
  - Only real exception is the use of simple ROOT (or ROOT+our data formats)
- Releases do change every few days, naming is as **CMSSW_X_Y_Z**. It is important since root files produced with one version can be not readable by other releases. Rule of thumb:
  - X: major release; no compatibility (usually) between release with differences in first digits
  - Y: medium differences expected. We used to give no compatibility guarantee, but as a matter of facts they could be ok. Check/ask on HyperNews
  - Z: maintain compatibility. **Indeed the advice is always to use, given X and Y, the release with highest Z at any moment**.

CMSSW releases

- You *cannot* use a random release with a dataset.
- DATA: Again, the best thing is to follow the recommendation in https://twiki.cern.ch/twiki/bin/viewauth/CMS/PdmV2012Analysis and similar

This dataset is ok with this release

- MC: Prep tells you which is the processing Release: use one compatible as explained in the previous page

Using the hint to use the release 5_3_Z with greatest Z, today it would mean to use 5_3_22 today

6. Find the correct version of the calibrations

- Calibrations are NOT coming with the release, for many reasons the most important is that they NEED the release to be constructed, and come later
- Correct Calibrations are ESSENTIAL if you want to reprocess reconstruction, but are important also if you are doing analysis (for example, they contain b-tagging efficiencies …)
- **Please note you need different calibrations for DATA and MC!**
- In CMS, calibrations come in the form of **G**lobal**T**ags (GT)

- The guide MUST be the staring point: here

Which Global Tags to use

- DATA: names are of type **GR_X_VYY(_ANZZ)**
  - **X** can be
    - **P**: used for prompt recostruction
    - **E**: used for express reconstruction (you do not need this)
    - **H**: used for HLT (you do not need this)
    - **R**: used for reprocessing
    - **_AN** in the end means it contains ANalysis level calibrations (i.e. btagging calibrations and similar)
  - Or **FT_X_YY**
    - These mean **FrozenTag**, used for official reprocessing

- RULE OF THUMB: if you are running on PromptReco, use the appropriate **GT_P_X_VYY_ANZZ** as in here
- If you use a Reprocessing, go with the **FT** as in here

Which global tags to use

- MC: easier than with data
  - **STARTXX_VYY**: samples created with realistic misalignment + calibrations
  - **IDEALXX_VYY**: ideal detector, perfectly aligned and with all calibration corrections =1
  - **MCXX_VYY**: like ideal but with broken channels as of 2010 implemented.
- **In most cases, you want to use the START GT**
- In general, take the one with correct XX (CMSSW release version), latest version
  - If using CMSSW_5_3_X, use START53_V29A
  - Full list available here

7. Run on it

- Here starts the fun …

- **Preliminary (ask your site admin / supervisor)**
  - Find a machine on which to work (which has access to CMS software + GRID tools) + some local disk
  - Create a CMSSW working area there (`scram projectCMSSW CMSSW_X_Y_Z`)
  - Correctly set the environment (`cmsenv`)
  - Have an analysis configuration which runs locally (`cmsRun yourConfig.py` works on a test local file)
    - You also have set the correct GT inside

  - **Now let's run it on the GRID**

CRAB(3)

- **CRAB** (CMS Remote Analysis Builder) is the tool which helps you to move from running on a local file to running on full datasets, irrespective where they are (well, almost …), using hundreds of CPUs at the same time

- **CRAB2** is the software we used up to … today
- If you need to invest time, please go directly to **CRAB3**, the software we are planning to use from no on!

- You need
  1. **A working CMSSW area**
  2. **A working local/remote CRAB3 installation**
  3. **… in which a working .py has been created and tested locally**
  4. **A list of datasets on which you want to run**
  5. **A GRID + CMS enabled certificate**
  6. **A place where to write the output**

- **1.-2. ask your system manager**
- **3. that is your analysis task!**
- **4. should be already clear from the discussion we had**
- **5. see later**
- **6. see later**

A GRID certificate. What for?

- When you run on the GRID, your executable will land on machines literally everywhere in the World (40-50 available sites)
- You cannot have standard login (**username/passwd**) on all of them; yet it is not admitted tasks do run without strict authentication
- Think of GRID certificate as a global substitute for username/password

- You need a two step process:
  1. Someone trustful guarantees your name is **Name Surname**
  2. CMS guarantees you are part of its manpower (== you can use the resources bought by CMS)

Certificates: 2 step process

1. You need to be authenticated by a trusted authority; most probably http://www.ncp.edu.pk/pk-grid-ca/ for you
   - Follow local instructions at the link
2. Once you have it, you must have it authorized by CMS: with a browser with the certificate onboard, visit here and apply

- **At this point, all CMS resources, worldwide, will accept you as a CMS user.**

Where to write your output?

- You have two options
  - Have it returned to your working area. It works if
    - The files are small (recommended<100 MB)
    - You have enough space in your working area (100 MB per job, 1000 jobs = 100 GB.. Beware!)
  - Save it to a Storage Element (~ GRID disk)
    - There usually you have much more disk available
    - These disks sit at Tier2s. To find one in which you are enabled to write you need to ask your **Country Resource Manager** where you have rights to write
      - The answer Should be in terms of a site like T2_PK_XXXX
    - If you are part of an analysis team , you could be granted access to a T2 where some of your collaborators sit
    - In most of the cases, better to write to a Storage Element …

So here we go … CRAB3

- (Full tutorial at https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookCRAB3Tutorial)
- To get in into your Path, just issue a
- `source /cvmfs/cms.cern.ch/crab3/crab.sh`
- You need to prepare a `crabConfig.py` per dataset you need to run
- The simplest one is very simple, and contains a few sections

crabConfig.py

6 main configuration areas, all with a plethora of possible options….

Section	Description
`General`	In this section, the user specifies generic parameters about the request (e.g. request name).
`JobType`	This section aims to contain all the parameters of the user job type and related configurables (e.g. CMSSW parameter-set configuration file, additional input files, etc.).
`Data`	This section contains all the parameters related to the data to be analyzed, including the splitting parameters.
`Site`	Grid site parameters are defined in this section, including the stage out information (e.g. stage out destination site, white/black lists, etc.).
`User`	This section is dedicated to all the information relative to the user (e.g. voms information).
`Debug`	For experts use only.

What to run?

```
config.section_("General")
config.General.requestName =
'tutorial_Data_analysis_test5'
config.General.workArea = 'crab_projects'

config.section_("JobType")
config.JobType.pluginName = 'Analysis'
config.JobType.psetName =
'pset_tutorial_analysis.py'
```

- Give a name to the task
- Give the .py you already tested locally

On which data/MC to run?

```
config.section_("Data")
config.Data.inputDataset =
'/SingleMu/Run2012B-13Jul2012-v1/AOD'
config.Data.dbsUrl = 'global'
config.Data.splitting = 'LumiBased'
config.Data.unitsPerJob = 20
config.Data.lumiMask = 'https://cms-
service-
dqm.web.cern.ch/…/PromptReco.txt'
config.Data.runRange = '193093-193999'
config.Data.outlfn =
'/store/user/<subdir>' # or
'/store/group/<subdir>'
config.Data.publication = True
config.Data.publishDataName =
'CRAB3_tutorial_Data_analysis_test5'
```

- Input dataset
- How many lumis per job?
- LumiMask
- On which Runs
- Which output file names to write (LFN)
- Make results available in DBS/DAS

Where to run (whose CPU use)?

```
config.section_('Site')
config.Site.blacklist = ['T2_IT_Legnaro']
config.Site.whitelist = ['T2_IT_Bari']
config.Site.storageSite = 'T2_IT_Legnaro'


…

config.Site.whitelist = ['T2_US*']
```

- Sites can be blacklisted (do not run there!)
- Sites can be whitelisted (Run only there!)
- **storageSite** is the final destination of your output

All US Tier2s

CRAB commands

In this section we provide a list with the currently available CRAB commands and their explanation. We will see how to use the commands as we go along in the tutorial.

Command	Description
`submit`	Submit a task.
`status`	Report the states of jobs in a task (and more).
`resubmit`	Resubmit the failed jobs in a task.
`report`	Get a task final report with the number of analyzed files, events and luminosity sections.
`kill`	Kill all jobs in a task.
`getoutput`	Retrieve the output files from a task.
`getlog`	Retrieve the log files from a task.
`uploadlog`	Uploads the crab log file to the CRAB cache in the server.
`checkwrite`	Check write permission into a site.
`checkHNname`	Check username extraction from SiteDB.
`purge`	Clean-up the user's directory in the schedd's and in CRAB cache.
`remake`	Recreate a CRAB project directory.

That's it …

`crab submit` launches the jobs, and in principle you should after some time (depending mostly on your jobs' duration) find the outputs in the chosen location (locally/SE)

`Crab status` gives you the instantaneous status at any moment… and dashboard helps you to monitor the situation: search yourself at

- http://dashb-cms-job-task.cern.ch/dashboard/request.py/taskmonitoring

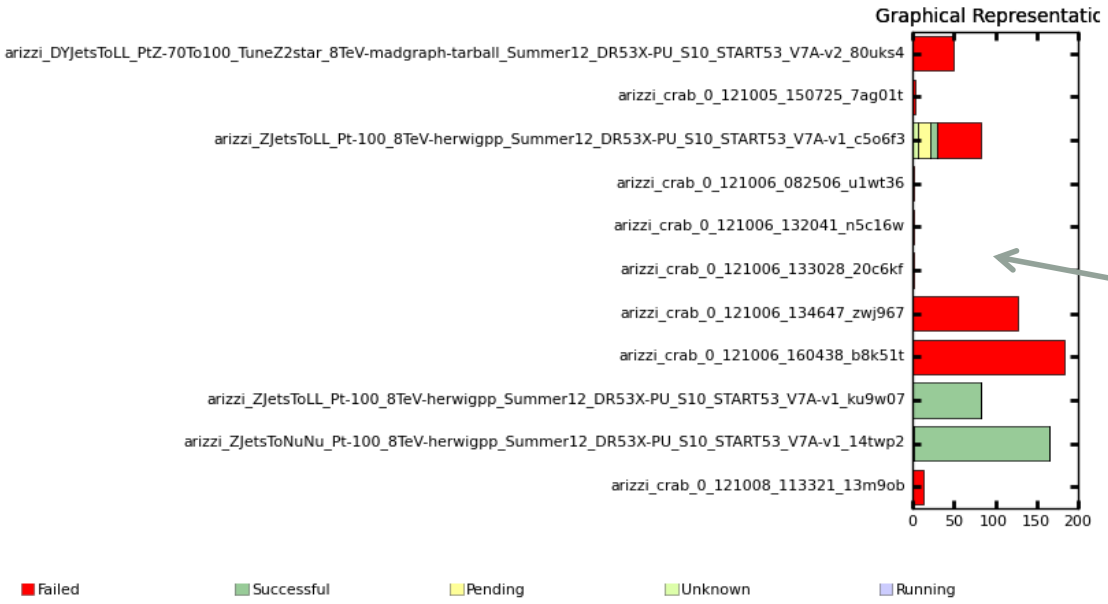
Select a User: Arizzi | Select a Time Range: Last 2 Weeks | Refresh: 5 Minutes | Help | User Support

Job Processing is not completed unless job GRID status is DONE. This page does not track further steps inside CRAB Server. Please ignore the GRID status for local submissions.

Your name

TaskMonitorId	Num of Jobs	Pending	Running	Appl Successful	Failed	Unknown	Completed Successfully	Consumed Time	Plots
arizzi_DYJetsToLL_PtZ-70To100_TuneZ2star_8TeV-madgraph-tarball_Summer12_DR53X-PU_S10_START53_V7A-v2_80uks4	49	0	0	0	49	0	0 out of 49	Time Info	Plot Selection
arizzi_crab_0_121005_150725_7ag01t	3	0	0	0	3	0	0 out of 3	Time Info	Plot Selection
arizzi_ZJetsToLL_Pt-100_8TeV-herwigpp_Summer12_DR53X-PU_S10_START53_V7A-v1_c5o6f3	83	16	0	7	54	6	0 out of 83	Time Info	Plot Selection
arizzi_crab_0_121006_082506_u1wt36	2	0	0	0	2	0	0 out of 2	Time Info	Plot Selection
arizzi_crab_0_121006_132041_n5c16w	1	0	0	0	1	0	0 out of 1	Time Info	Plot Selection
arizzi_crab_0_121006_133028_20c6kf	1	0	0	0	1	0	0 out of 1	Time Info	Plot Selection
arizzi_crab_0_121006_134647_zwj967	127	0	0	0	127	0	0 out of 127	Time Info	Plot Selection
arizzi_crab_0_121006_160438_b8k51t	183	0	0	0	183	0	0 out of 183	Time Info	Plot Selection
arizzi_ZJetsToLL_Pt-100_8TeV-herwigpp_Summer12_DR53X-PU_S10_START53_V7A-v1_ku9w07	83	0	0	82	1	0	0 out of 83	Time Info	Plot Selection
arizzi_ZJetsToNuNu_Pt-100_8TeV-herwigpp_Summer12_DR53X-PU_S10_START53_V7A-v1_14twp2	166	0	0	164	1	1	163 out of 166	Time Info	Plot Selection
arizzi_crab_0_121008_113321_13m9ob	14	0	0	0	14	0	0 out of 14	Time Info	Plot Selection
Sum Total	712	16	0	253	436	7	-	-	-

Your tasks



Uhm, not too well!!!

Select a User: Arizzi Select a Time Range: Last 2 Weeks Refresh: 5 Minutes Help User Support

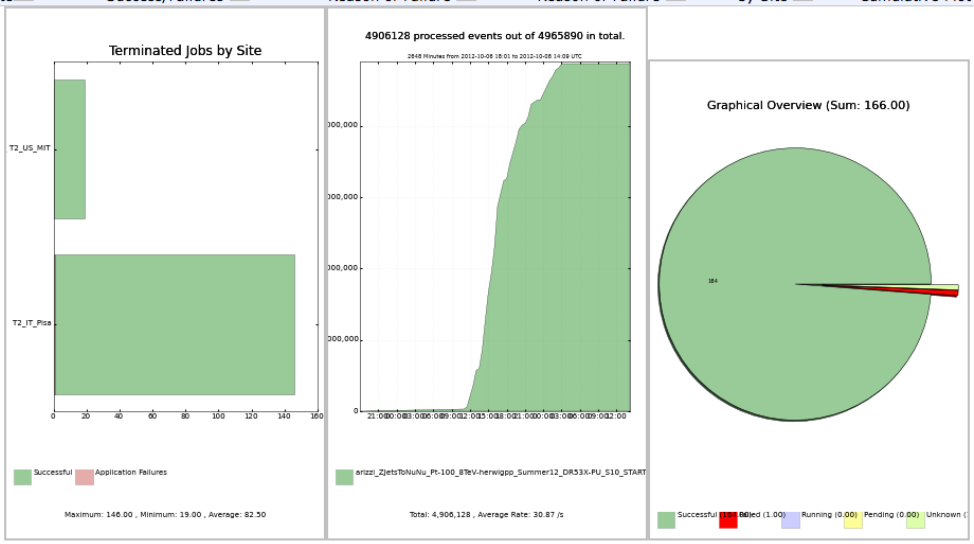
Task: arizzi_ZJetsToNuNu_Pt-100_8TeV-herwigpp_Summer12_DR53X-PU_S10_START53_V7A-v1_14twp2 All Jobs Back to all Tasks This Task

Job Processing is not completed unless job GRID status is DONE. This page does not track further steps inside CRAB Server. Please ignore the GRID status for local submissions.

TaskMonitorId	Num of Jobs	Pending	Running	Appl Successful	Failed	Unknown	Completed Successfully	Consumed Time	Plots
arizzi_ZJetsToNuNu_Pt-100_8TeV-herwigpp_Summer12_DR53X-PU_S10_START53_V7A-v1_14twp2	166	0	0	164	1	1	163 out of 166	Time Info	Plot Selection

Dashboard Plots

Successful Jobs Distributed by Site | Processed Events by Site | Terminated Jobs in terms of Success/Failures | Application Failed Jobs by Reason of Failure | Grid Aborted Jobs by Reason of Failure | Terminated Jobs by Site | Processed Events Cumulative Plot | Terminated Jobs Distributed over Time



Click on a plot to increase its size.

SchedulerJobId	Id in Task	Appl Status	Appl Exit Code	Grid End Status	Retries	Site	Submitted	Started	Finished
https://wms006.cnaf.infn.it:9000/jV7Q7sn7fKhQOA5L4jmeUw	1	Appl Succeeded	0	Done	2	T2_US_MIT	2012-10-07 07:56:31	2012-10-07 08:10:23	2012-10-07 21:48:56
https://wms019.cnaf.infn.it:9000/NDUPJ243Ri3Ep4oJJ-AHbw	2	Appl Succeeded	0	Done	1	T2_IT_Pisa	2012-10-06 18:59:29	2012-10-06 20:36:06	2012-10-07 14:59:33
https://wms019.cnaf.infn.it:9000/VhkiaPt24ZLAbtKQDmq13Q	3	Appl Succeeded	0	Done	1	T2_IT_Pisa	2012-10-06 18:59:29	2012-10-06 20:35:17	2012-10-07 17:49:55
https://wms019.cnaf.infn.it:9000/ih85OqJ4iBwqBNGLblKegw	4	Appl Succeeded	0	Unknown	1	T2_IT_Pisa	2012-10-06 18:59:29	2012-10-06 19:07:11	2012-10-08 15:07:55
https://wms019.cnaf.infn.it:9000/ibDrb33S5sMvvIDiDSqPMQ	5	Appl Succeeded	0	Done	1	T2_IT_Pisa	2012-10-06 18:59:29	2012-10-06 19:01:31	2012-10-07 16:48:12
https://wms019.cnaf.infn.it:9000/e2AyT54MYAFdhvM6-2-OpA	6	Appl Succeeded	0	Done	1	T2_IT_Pisa	2012-10-06 18:59:29	2012-10-06 20:04:25	2012-10-07 14:15:56
https://wms006.cnaf.infn.it:9000/rNoBebR2PNPvT6bV6IBRtQ	7	Appl Succeeded	0	Done	2	T2_US_MIT	2012-10-07 07:56:31	2012-10-07 08:10:47	2012-10-08 00:51:42
https://wms019.cnaf.infn.it:9000/1Cu_vNvnuyL2c855J1J5aA	8	Appl Succeeded	0	Done	1	T2_IT_Pisa	2012-10-06 18:59:29	2012-10-06 19:01:56	2012-10-07 16:06:56
https://wms019.cnaf.infn.it:9000/BqCbclLWIEWYi4aGqgdETw	9	Appl Succeeded	0	Done	1	T2_IT_Pisa	2012-10-06 18:59:29	2012-10-06 20:38:54	2012-10-07 18:24:59
https://wms019.cnaf.infn.it:9000/AGLD_iTb657CZ7oMVK7UGw	10	Appl Succeeded	0	Done	1	T2_IT_Pisa	2012-10-06 18:59:29	2012-10-06 19:11:50	2012-10-07 13:27:38
https://wms019.cnaf.infn.it:9000/i5DzBhEw3ynoQCQez4ELfQ	11	Appl Succeeded	0	Done	1	T2_IT_Pisa	2012-10-06 18:59:29	2012-10-06 19:11:04	2012-10-07 14:39:10
https://wms019.cnaf.infn.it:9000/6gcu7UTPZEkGJyhIpc2iOA	12	Appl	0	Done	1	T2_IT_Pisa	2012-10-06	2012-10-06	2012-10-07

Clicking on a task, you can see where it went, and how successful it was

Most of the jobs went to site T2_IT_Pisa

What if something got wrong?

- If all the jobs fail, it is most probably YOUR fault
  - **Does your configuration crash?**
    - Test it better, on more than 10 events
  - **Does it leak memory?**
    - Look with “top” at memory usage over at least 100 events, locally
  - **Is the output file VERY large (> 500 MB)?**
    - Use fewer LSs per job

- If some failed, there should be automatic resubmit; you can still manually resubmit later for a very last attempt
    - `crab resubmit`

After all this….

- You should have in an “accessible place” all the files you need
- (or, you can get them locally via `crab getoutput`)

- So you have run on data and MonteCarlo… what next?
- This is very analysis dependent, and usually boils down to running some ROOT macros on the results and get some nice plots.

(some) conclusions

- We are not trying to sell that doing analysis @ CMS is easy – but indeed with some tools we have evidence it can be done (>300 papers published by now)

- You will definitely need help at the beginning; please consider
  - WorkBook and Software Guide
  - Hypernews: don't be shy … there are literally hundreds of people willing to help
  - Still, identify your local expert … asking someone close to you can let you save days!

- **Enjoy Physics @ CMS!**