

Skimming and HLT objects

J. Incandela

Oct. 24, 2007

Joint Physics & Trigger Week



Foreword

- The 2007 analyses, CSA07, skims, etc. have raised a lot of important issues
 - A complicated situation that cannot be completely factorized
 - It's clear that we will need to give serious thought to all issues:
 - Contents and sizes of data-tiers
 - (see for instance Shahram's talk yesterday)
 - How to manage data distribution
 - (see for instance any of Ian's 23 CSA07 talks)
- We will review all of this in the coming months in order to come up with a better plan
 - One other issue that arose in the CSA implementation of the computing model is the question of how persistent (i.e. trigger) information is used in defining datasets.



Access and Efficiency

- There are various possible philosophies one could use for the basis of a computing model
 - Brute force:
 - Infinite resources
 - CPU, Storage, Transfer speeds, etc. adequate to allow everyone in CMS to run on any datasets of any size whenever they want
 - Finite resources for “elite” physics users/groups
 - Limited resources accessed only by the “really good” people
 - » This solution is unstable and evolves to the modus operandi of the infinite resources case (since we know we are all “really good”) but with inadequate resources to make it work
 - Intelligent design: Affluence with finite resources
 - Select the smallest datasets that maximize physics impact and flexibility in spite of finite resources
 - Make it possible for everyone on CMS to do physics without pain
 - This is the basic intent of the CMS Computing Model (I think)



Computing Model

- Let's assume that a fundamental goal of the CMS computing model is easy access to well understood data samples that enable physics to be done efficiently and effectively by all members of CMS
 - Trigger, Primary Datasets (PD) and skims all play a role
- Computing TDR (\S 2.4 Data Flow Overview)
 - “The CMS DAQ system writes DAQ-RAW events (1.5MB) to the High Level Trigger (HLT) farm input buffer. The HLT farm writes RAW events (1.5 MB) at a rate of 150 Hz. RAW events are classified in $\mathcal{O}(50)$ primary datasets depending on their trigger history (with a predicted overlap of less than 10%). Primary dataset definition is immutable. An additional express-line is also written with events that will be reconstructed with high priority. Primary datasets are grouped into $\mathcal{O}(10)$ online streams in order to optimise their transfer to the offline farm and the following reconstruction process. “



First Iteration

- **CSA07**
 - 6 (+1) Primary Datasets (PD) – much less than the $\mathcal{O}(50)$
 - Driven by physics & constraints from Computing and Offline.
 - **Keep the number of workflows manageable**
 - Skims were to run on ~1 PD
 - Physics groups held to ~3-5 Skims each
 - **For physics, fewer skims means they be more inclusive.**
 - They also widely make use of volatile RECO/AOD information
- **Observations**
 - The PD are not unlike what one would call Online Streams in terms of their content and sizes (if not functionality)
 - The $\mathcal{O}(50)$ physics skims are as numerous as the CTDR plan for PD's (but differ from PDs in a very important way).



Persistence

- The skims lack a fundamental aspect of PDs
 - They are not persistent – i.e. the skims are not based upon intrinsic event qualities
- Trigger information is intrinsic
 - It is what defines the data we collect.
 - It stays with the event and so makes an event's origin traceable, and can be used to reconstruct any dataset, provided the events are not lost or destroyed.
- Just stating the obvious: but this is important
 - We get to keep only a few out of every 100k collisions.
 - Triggers are the single biggest factor in the determination of our datasets
 - (assuming the machine & detector are working ...)



A use-case

- Having easy access to persistent datasets allows one to track important changes in your analysis
 - Example:
 - A discovery involves ~ 25 events in a peak over a background of ~ 5 events
 - New alignment constants are released and $\sim 15\%$ of the events change in the signal peak.
 - If the event sample under study is a skim that's not based on persistent information, the events that dropped out of the signal peak may not be in the skim sample anymore!
 - Must go to the PD and re-run a “new” skim for the complement of the set of events you have.
 - This will not be easy if the PD is large, and spread over T1's which are themselves not easy to access.



Future Iterations

- Skimming is one of the hardest parts of CSA07.
 - If this is not overcome, we'll need to adjust the model
 - If it is overcome, we may still want to adjust the model
- Some obvious questions (for me at least):
 - Can we recover the plan for $\mathcal{O}(50)$ PD?
 - This would solve a weakness of the CSA07 model
 - With 6 PD, I predict that it is not going to be practical to go back to the PD as often as we like and yet these will be the only persistent datasets
 - We would have $\mathcal{O}(50)$ persistent PD
 - Skims could be more finely tailored and run more easily (since they would run on much smaller samples) and we might even copy some PDs closer to home (i.e. to T2s) or make sub-PDs for this purpose
 - If not, then I believe that the skims must absorb this role to the largest extent possible
 - It could be an issue of nomenclature only. What we now call PDs could be streams, what we now call skims could be replaced by a T1 based effort to construct 50 persistent PD



Data placement

- CSA case
 - PD (persistent) large and only available in complete form (all data-tiers, all events) at T1
 - Skims (not persistent) available at T2
- Start of run (and probably even later...)
 - It is not clear that it will be easy to go back to datasets that exist in full glory only on T1's
 - So, in addition to AODs for larger samples (in numbers of events), it would be good to have persistent datasets of *manageable sizes* with full event information at T2's
 - Allows flexibility and speed
 - Easy access for all members of the collaboration to allow us to quickly find and understand problems and to test solutions more thoroughly before committing to a re-reconstruction of all data.
 - Allows one to define non-persistent datasets
 - Changes can be traced without need to go back to T1 (just return to the T2 persistent sets from which it originates)
 - Allows more local control and less stress on T1s



Other possible benefits

- Using persistent (i.e. Trigger) information to define datasets has another benefit
 - Broader involvement in trigger definition and understanding of trigger performance.
 - The triggers are not cut in stone and would evolve more quickly if they are used to define datasets as far as possible



Summary

- So the basic message is that we should consider
 - Taking the idea of persistent definition of datasets as far as we can reasonably go. This has the potential for greater efficiency overall by allowing
 - broader access to traceable information
 - broader participation in trigger definitions
 - faster and more thoroughly tested solutions to problems
 - At some point we have to break with the trigger-based selection
 - This should be after the point where full-service datasets are relatively small and easy to access (already at T2 ?)
 - To allow more local control of final dataset definition
 - NB: Very large datasets cannot be avoided in some cases but may not need to be accessed frequently or by many people and maybe could be run at the T1's

Additional info



Ex. From CDF

- Discussing these issues with CDF people I got the following comment
“On CDF we're rigorous about matching offline objects to trigger objects... this is the only way to accurately account for your trigger efficiencies. Assume [an event] passes the muon trigger, but for some reason failed the electron trigger. Now offline, if the muon fails our ID requirements but the electron satisfies them, it might be a $t\bar{t} \rightarrow e\nu b\bar{b} jj$ candidate. Understanding the trigger efficiency for such an event would be very difficult. In our case, if a dilepton veto didn't kill it, the fact that the primary electron in the event doesn't match to any trigger object would kill it as well.”
- Of course you must study triggers at all times and all luminosities to make sure that they are optimal. But then you have to accept them.