Data Processing Status Update

Stefano Lacaprara & Marco Milesi

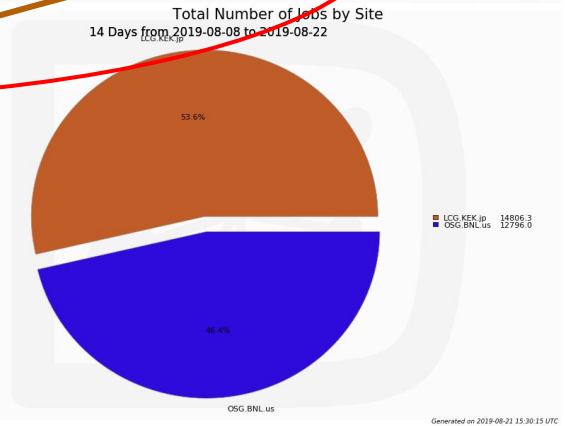
Data Production Meeting 22/08/2019

Bucket 7 Status

- **Production of mdst/cdst for hlt_skim**
- **Production of mdst (only) for All events**
  - https://agira.desy.de/browse/BIIDP-1643
  - Release-03-02-04
  - **GT: online + data_reprocessing_prompt_bucket7**
- **Status: DONE**
  - **359** runs
  - **10434** jobs total for hlt_skim
  - **56092** for all events
    - Few runs failed for problem with input files:
      - Error in <TFile::Init>: /group/belle2/dataprod/Data/Raw/e0008/r01925/sub00/physics.0008.01925.HLT3.f00010.root not a ROOT file
      - 1920, 1925, 1932, 1935, 1936, 1940, 1941, 1942, 1943, 1946, 1951
    - Not moved yet to final path
- Output path:
  - **/group/belle2/dataprod/Data/release-03-02-04/DB00000677/bucket7/e0008**

Proc 9 on Grid: Status

- Some activity!
  - https://agira.desy.de/browse/BIIDP-1587
  - Hideki-san fixed processing of long run with more than 1000 input files:
    - Thanks!
    - https://agira.desy.de/browse/BIIDCD-888
- **ProdID 8521 still running** (exp 8)
    - RawProcessing still NOT done:
    - Since two weeks ago (8/8/2019) [**46076** jobs Waiting]
      - **~12800** Done at BNL (in one day)
      - **~15000** Done at KEKCC (in 14 days)
    - Now job draining at KEKCC (downtime 20-26/8)
  - **29946 waiting (22 Merge waiting)**
  - **Exp3 DONE, Exp7 1 trans waiting (>1000 input files, fixed)**
- I'd like to process Bucket7 on the grid
  - **ALL at BNL, if possible! How?**
  - All files are already staged



Good, Bad (, and Ugly) Runs

- Proc9 https://agira.desy.de/browse/BIIDP-1712
- Bucket7 https://agira.desy.de/browse/BIIDP-1746
- How to provide good/bad runs to users:
  a. **Restrict access to bad runs (setfacl)**
     - Error safe for users
     - What if you want to look at bad runs specifically?
  b. **Different path for bad runs: eg**
     - `.../bucket7/e0008`
     - `.../bucket7/e0008/BadRuns/`
       - Should be safe also for users
  c. **Or in run directory name:**
     - `...bucket7/e0008/r01917_bad`
       - Was done for early processing for exp3.
       - Not very safe is user `glob.glob("bucket7/e0008/r0*")`
  d. **Or for both bad and good:**
     - `.../bucket7/e0008` for all runs
     - **sym link** to `.../bucket7/e0008/(GoodRuns|BadRuns|UglyRuns)/`
- **I prefer solution b) or d).**
  a. All solution requires today some manual work by DataProcessing

Good/Bad Runs: longer term

- No solution based on path on filesystem is a long term one.
  - Eventually mdst will be on grid (only).
    - It is not easy (or possible) to change path.
    - We could change the metadata for each file (eg “bad” for bad runs), but it would require non negligible work.
- A better solution should involve basf2 framework
  - Feed basf2 (and/or gbasf2) with a centrally-blessed list of good runs
    - There could be more than one
      - Eg: If you are not using muons and/or KL, KLM problems do not affect your analysis
  - Framework skip bad runs according to list.
- It could be a payload in DB, that would allow easy distribution of updates
  - Default is run on good runs only, unless user explicitly ask otherwise.

Online vs Offline # events in run

- On/Offline lumi mismatch
  - Reported by Xing-Yu Zhou
  - https://agira.desy.de/browse/BIIDP-1734
- Detailed investigation of all Exp8 runs
  - Comparing # events in elog (online)
  - And offline
- Results in plot
  - Difference (online-offline)
  - Ratio (online/offline)
    - Vs Run number
    - Distribution (log scale)
- Outliers are non negligible and large!
  - Not sure I can trust the online from elog.
- Still no conclusion
  - Don't know what to check