BRAHMS Offline Data management

Life cycle of event data

In order to discuss how we keep track and manage data/data files I will attempt to describe how data files or part of files will progress and be used in analysis. Such document/information should involve into specifications for reconstruction information database definitions, what the tools should do for us etc.

- Raw Data Generation
- Calibration Data Generation and usage
- Reconstruction pass
- Particle ID, DST (RDO) generations
- PhD /microdst/Tree analysis (Physics analysis)

Raw data

The raw data are collected on the SUN 6cpu 3500 in the spool area. A given run i.e. same set of physics, trigger conditions will be store in individual sequence files named Runxxxxxseqyyy.dat. Each file should be approximately .5-1Gb in size. Smaller sizes will result in access penalties for HPSS.

- The files are stored in the spool area on opus.
- The files are managed by a script (xxx) that sends the files to the RCF HPSS file system. To reduce file sizes and increase efficiency of subsequent analysis we should seriously consider converting DAQ raw format to ROOT files. This would incure no loss of information (except for possible no-hit channels in TOF,H1,H2,…)
- Files are available for the users on the ONCS compute farm until deleted. The deletion process uses by following rules
    - Oldest files deleted first. Files must be saved on HPSS tape! This requires an API that can inquire if this has been done.
    - Files can be marked as *special* and not deleted until .. this could e.g. be calibration files. Is this possible at present

Calibration pass

With the large amount of data that will be generated in a typical year of RHIC running it will be quite difficult both due to computer resources (RCF bandwidth) and human resources (QA) to read/analyze all data files **after** they have been recorded in HPSS. It is essential that monitoring and a first calibration takes place during each run to record status of detectors, bad channels and other simple tasks that are used to determine the quality of the data. Such information can also be used to select what runs should be used to determine calibration constants.

- Read selected set of data files from HPSS, analyze in the crs farm; The output is histograms, data files/ information to be inserted into the calibration data bases.
- Some detectors requires additional information before calibrations can be done. As examples
  - Tracking in MTPC1 to get vertex infor for BB calibrations
  - Global tracking to select pion's for TOF calibrations.
- Thus for some data 2 data sets may be read and partially reconstructed.

Reconstruction Pass

The raw data, and the calibrations conditions database is used to generate reconstructed tracks both in FS and in MRS, as well as RDO for global detectors. Most likely only a preliminary PID is generated since this requires tracks, and well calibrated time-of-flight. The output is a second set of data files. These do *not* include all of the original information. The sizes are probably still such that a run is in the order of .5-1 Gb (1:4 reduction over raw data). Should they become much smaller data should be merged.

Particle identification and DST generation.

The RDO outputs is used to generate PhDs (DST) from the reconstruction pass, as well as a better PID. The output again is (reduced) compared to the RDO, and should be well suited to physics analysis. These files may well be much smaller than the RDO's. In that case should multiple files be merged?
This analysis will take place on the CAS nodes, and at institutions.

MicroDST analysis

Keeping track of the files.

We will need a database that contains for each steps quite a bit of information. The information required should be sufficient to re-do the reconstruction task. It should (indirectly) point to run conditions. At the DST level there may have been other data selections involved example wise only take data with 1/2 tracks in FS. This later comments point to need for a kind of tag database either per run or per.

- Input file(s)
- Output file name, size, location
- What standard code was used - version.(also version of Brat libraries!! Even possible ROOT)
- Date, execution time, machine run on
- Account used.
- Scripts used? - As a way to determine parameters applied.
- Physics selection criteria.