Event Overlaps

Jack Cranshaw

The Issue Described

- In Run 2 we run trains of derivations which turn reconstructed AOD (physics_Main) into derived AOD (~80 streams).
- If these overlap significantly, then this has a magnified effect on resources.
  - We need to be able to monitor this at the stream-stream level, i.e. 2-stream overlap.
  - This cannot be done during processing because there are roughly 10 trains, so no process sees all the output streams.
  - Primarily needed when new releases/caches are cut (1-2 times per month) as fractions should be relatively constant for given trigger configurations, BUT I wouldn't be surprised if they run it on a fixed sample in a nightly at some point.
  - This needs to be presented to the managers in a matrix format. They will decide if there is a problem and follow-up with the appropriate people. When a change is made, then the developers may want to re-run the overlap check by hand.
  - The basic query is well fitted to an SQL type query doing a join on the run-event primary key.

Status

- First there was time, but no data; then there was data, but no time.
- Initially tried Pig as it has a close to SQL interface. Had issues.
- Other possibilities
  - Java user function with Pig
  - Actual map reduce job (How to access multiple datasets in a job or build serial job?)
  - Is this data in Hbase?