Outcome of the Software Week 16--20 June 2008

Data Quality parameters

Back to the initial proposal of a finite set of quality parameters combinations, connected to the *files* table via a QualityID.
The BK will provide an interface for the data quality team to update the values of the quality table.
About how the quality is computed, there was no definitive answer. It seems that, until real data are there and people will start to work with them, it is not easy to figure out the exact definition of data quality. For sure, the quality should not depend on the status of some subdetector. In fact this information should be rather stored in the *data taking conditions* table. Then, it seems that the data quality should be something computed at data analysis level. Then, maybe, all the files coming from the same production could have the same data quality flag, thus we could associate the data quality flag to the production rather than to the file.

The concept of production

Maybe it is a good idea to create a *production* table, because there are many quantities which are related to the production, rather to the single file. We would link it to the *jobs* table via the production number. And in the production table we could store: program name and version, processing pass, simulation condition Id (or data taking condition Id for real data) and the quality flag. As said in the previous utem, from what Olivier said in the meeting, it seems possible that the data quality is set on a production basis.

The concept of production for raw data

At some point it was proposed to identify the production with the LHC fill number (at each filling of the machine with protons, this number increments). Corollary: what is the "processing pass" of RAW data? Can one query RAW data on a "production" (i.e. fill number) basis?

As long as a new run is started when a new fill starts, it is possible to identify the concept of production with an LHC fill. And , in the same way, data of the same fill have all the same data taking conditions. As long as this is true, the identification of a production with a fill is ok. So the hierarchy to group data is: run (the smallest unit), then fill, and finally a set of fills with same data taking condition Id. In this way, yes, it is possible to query data on the basis of the fill.
And about the meaning of Processing Pass for real data, this is a set of data which have been processed with the same version of the application, in this case Moore.

Data Taking Conditions

Previously called *data taking period* has been renamed to *data taking conditions* because it is not related to a time period.

The *generator* has been moved to the *simulation conditions* table.

-- ElisaLanciotti - 20 Jun 2008

This topic: LHCb > Swjune08
Topic revision: r2 - 2008-06-20 - ElisaLanciotti