Bookkeeping issues

When looking into how to use the current schema for online data, I came to severe limitations (largely due to shortcuts) in the current schema, which could even affect any data type if even minor changes are done in some conventions. This should by all means be avoided in such a Database and hence the issue should be addressed as soon as possible.

Job referencing: missing unique name

The current schema contains as job information in the table */jobs*: CONFIGNAME, CONFIGVERSION, JOBDATE and JOBID. An additional table */jobParams* is related to */jobs* via the JOBID and contains many parameters related to the job. One can immediately notice that the only parameter that characterise uniquely a job in */jobs* is JOBID, which is an automatically generated identifier. A job doesn't have a "name" that is meant at being unique such that a query can be made on it. There a NAME parameter though in */jobParams*. Is is supposed to be unique?

Looking at the web page, one could imagine such information exists as there is a "Job Lookup" form. Unfortunately this query relies on the file naming convention:

```
<file_name> = <job_name>_<step>.<extension>
<job_name> = <production>_<job_in_prod>
```

Similarly the "production lookup" relies on the above convention. For example the FETC in stripping jobs cannot be found with these queries... For example querying production 2000 uses the following URL:

```
http://volhcb05.cern.ch:8080/BkkServlets/DataSets?prev_fname=-&page_number=1&sql_user=%20f.filena
```

Even job queries (from the jobId link of a file is made through the file!). For example for job 12109574 the URL is

```
http://lhcbbk.cern.ch:8080/BkkServletsWrite/Select?job=/lhcb/production/DC06/phys-v2-lumi5-BcVegP
```

The primary cause of this limitation is that the job in */jobs* doesn't have a NAME as a key. If we are to identify a "run" to a "job", one has to be able to query a run by its number without having to rely on naming conventions of the files! This is as well very bad for production as we have seen above. A job should have a NAME that is unique... Action: BK developers

File insertion to an existing job

Currently, although files are the primary source of queries, they are inserted as an "outputfile" of jobs as shown in the following excerpt of an XML bookkeeping file:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE Job SYSTEM "book.dtd">
<Job ConfigName="DC06" ConfigVersion="Stripping-v31-lumi2" Date="2007-11-28" Time="11:21">
  <TypedParameter Name="Production" Value="00002000" Type="Info"/>
  <TypedParameter Name="Job" Value="00000361" Type="Info"/>
  <TypedParameter Name="Name" Value="00002000_00000361_3" Type="Info"/>
  <TypedParameter Name="Location" Value="LCG.CERN.ch" Type="Info"/>
..........
  <TypedParameter Name="ExecTime" Value="17546.2892981" Type="Info"/>
  <InputFile    Name="/lhcb/production/DC06/v1r0/00002000/FETC/0000/ETC_00002000_00000361.root"/>
..........
  <OutputFile   Name="/lhcb/production/DC06/v1r0/00002000/SETC/0000/SETC_00002000_00000361_3.root
    <Parameter  Name="EventType"     Value="10000000"/>
    <Parameter  Name="EventStat"       Value="1825"/>
    <Parameter  Name="Size"        Value="19939"/>
    <Quality Group="Production Manager" Flag="Not Checked"/>
    <Parameter  Name="MD5SUM"        Value="548417be0c97b535eced95b57bd1f6ce"/>
    <Parameter  Name="GUID"        Value="3A44ED5E-739D-DC11-BEAC-000E0C4D35D9"/>
  </OutputFile>
  <OutputFile   Name="/lhcb/production/DC06/v1r0/00002000/DST/0000/00002000_00000361_3.dst" TypeN
    <Parameter  Name="EventType"     Value="10000000"/>
    <Parameter  Name="EventStat"       Value="1825"/>
    <Parameter  Name="Size"        Value="1128071349"/>
    <Quality Group="Production Manager" Flag="Not Checked"/>
    <Parameter  Name="MD5SUM"        Value="38fc9853ad1a0530e4871c0763abdad2"/>
    <Parameter  Name="GUID"        Value="6821205D-739D-DC11-BEAC-000E0C4D35D9"/>
  </OutputFile>
...........
```

Due to the limitation explained above with *job* referencing, it seems unlikely that a file can be added a-posteriori to an existing job. *This should however be verified with the XML DDL and the BK registration code*. Action: Marianne

What can also be noticed from the above XML corresponding to stripping job is that the file parameter *EventInInputStat* is not set for neitehr the DST nor the SETC. It should be equal to the number of events in the FETC. This is useful information for physicists who use stripped data, as they have to knwo from how many original events this corresponds to. Action: Joël

Missing file and job parameters

Although the current schema may be adequate for simulation, there are a few additional parameters that are necessary for real data.

- Job/run parameters
    - Second time-stamp: a run should have information about its start and end time. Hence one additional time-stamp is needed. Should be part of the */jobs* table.

- File parameters
    - Strangely enough there is no time-stamp for files, neither in the */files* nor in the */fileParams* table. As for jobs, files should have two time-stamps in the */files* table. For MC data, they can be identical and equal to the job time-stamp.

Useless or unused job parameters

It seems some parameters are either redundant or useless in the */jobParams* table.

- *<application>* and *<application>_Version* (where <application> is Gauss, Boole, Brunel, DaVinci): this is very much MC-specific and redundant with the information in *ProgramName* and *ProgramVersion*
- *JOB_IDENTIFIER*, *PREFERRED_SITE*, *PRODUCTION_IDENTIFIER* don't seem to be used and are probably useless
- *Job* is irrelevant
- *SUBMIT_DATE* and *SUBMIT_TIME*: not used, should be used, but why not a time-stamp?
- *StatisticsRequested*: unused, should it be?

-- PhilippeCharpentier - 14 Dec 2007

This topic: LHCb > BKLimitations
Topic revision: r3 - 2008-01-07 - PhilippeCharpentier