Distributed Databases in LS1 and Run2

Dario Barberis

Genoa University/INFN

Planning DB activities in LS1+Run2

- The long LS1 period gives us the possibility to re-think, at least partially, our database usage and deployment and data access models in view of the start of Run2 in 2015
- CERN is re-negotiating Oracle licences and will most likely not cover any longer the costs of Oracle licences at Tier-1s
- A lot of experience has been gained in 4 years of operations with real data and it is a good time to revise our infrastructure
- Today we discussed only the plans for distributed database deployment
  - Current plans for database software developments and for the deployment of non-distributed systems are assembled in https://twiki.cern.ch/twiki/bin/viewauth/Atlas/DatabasesLS1Planning

Today's discussions

- Medium-term planning by the CERN IT-DB group
  - ■ Oracle licence re-negotiations
- Deployment and access model for Geometry, Trigger and Conditions DBs
    - ➢ These are the databases accessed by interactive/batch/Grid jobs
  - ■ Oracle+Frontier servers
  - ■ DB Releases
  - ■ Conditions data files in CVMFS
- Evolution of TagDB and its deployment model
- Muon Calibration Centres at Tier-2s
- Not discussed today:
  - ■ AMI – replicated from IN2P3-CC to CERN
  - ■ COMA – distributed with TagDB to support its functionality
  - ■ Access to metadata databases from Grid jobs (is it needed?)

Proposed plan for Oracle+Frontier

- The current system with a few sites having Oracle+Frontier basically works
- At the end of October I asked the 5 Tier-1s if they intend to continue providing an Oracle database for conditions data and a Frontier server for Run2 (till end 2018).
- 4 Tier-1s (BNL, IN2P3, RAL, TRIUMF) have no problem in continuing with the present system; KIT would like to phase out Oracle.
  - We can stay with 4 servers (2 in Europe and 2 in America) plus CERN.
  - We can keep this set-up till the end of Run2.
- In practice:
  - Don't touch anything till the end of Run1 and its intensive analysis (Summer 2013)
  - In the meantime, evaluate how to share the sites that now use KIT as primary or secondary server
  - Summer 2013: edit site assignments in AGIS, wait 1-2 weeks, check that there are no connection attempts to KIT, switch it off.

Database Releases (1)

- DB releases contain a snapshot of the Conditions, Trigger and Geometry databases that is used to generate and reconstruct simulated events
  - DB releases served us very well for many years for production and analysis tasks
- DB releases are incremental: the latest release can always be used with all software releases and older data
- For real data processing, DB releases are no longer sufficient as some of the calibration constants are in "conditions data files" referred to by the COOL database
  - Conditions data files are distributed to all sites and placed in HOTDISK
  - Now available also through CVMFS
- For reprocessing, special "Conditions DB Releases" are built that contain, in addition to the base DB release, the conditions data files that are used for a given run range

Database Releases (2)

- The "push model" with DB releases, conditions data files, CDB releases and HOTDISK at each site was designed before more modern web-based technologies were available
  - And worked OK till now
- This model doesn't scale as the accumulated data keeps increasing
- A "pull model" based on caches with Frontier and CVMFS looks more promising:
  - Frontier gives access to database information
  - CVMFS gives access to software and conditions data files
- Indeed already now sites with CVMFS should see no access to their HOTDISK – to be checked

Plan for DB releases etc.

- Remove HOTDISK from all sites that have CVMFS
  - May have to keep it for a while at Tier-1s as transit stop
  - In any case we ask all sites to install CVMFS…
- Stop building Conditions DB releases and take conditions files from CVMFS for reprocessing
  - This was the recommendation for the recent campaign but it was not followed
- Move progressively all applications to the use of Frontier instead of DB releases
  - Tests still missing for the Trigger MC database; all the rest is OK.
- Stop building DB releases at the end of Run1.
  - Keep the DB release technology for stable, large-scale production and reprocessing campaigns
- Last but not least: convince the detector communities to put their calibration data in COOL instead of the conditions data files

TAG Sites & Sizes (Dec 2012)

Site	Data type	Used space	Free space	Total space
CERN[(*)]	All real data	11	7	18
TRIUMF	Real data >2009	16	29	45
DESY	MC 09/10 (now MC12)	8	7	15

[(*)] Data at CERN are replicated 2 times

- MC12 starts being uploaded to DESY
- Data sizes for most collections can be reduced up to a factor 2 by reducing the indices
  - This explains the differences between CERN and TRIUMF
- As both TRIUMF and DESY can keep their instances going for the next 12-18 months (THANKS!) we can hopefully in the meantime develop and deploy a new back-end (see later slides)

DB's Tag simplification proposal

- Let's separate the 2 functions: DPD and EventIndex
  - Files and database don't need to have the same contents nor the same storage technology
- Oracle storage seems not to be the best choice for this kind of information
  - Schemas with no flexibility, cost issues
- Modern structured storage technologies are much cheaper, easier to use and faster for sparse searches over large amounts of data
  - CERN is going to set up a Hadoop service following the Database TEG report last Winter
  - Data can be stored in Hadoop/HBase as key-value pairs
    - One record /event, keyed by RunNo+EventNo+TriggerStream
    - Every processing cycle adds information to the event records:
      - First info provided by online: trigger pattern, lumi block, RAW pointer
      - Pointers to reconstructed data and physics info (if needed) added by Tier-0 and every reprocessing step

Tag Evolution Proposal+Plans

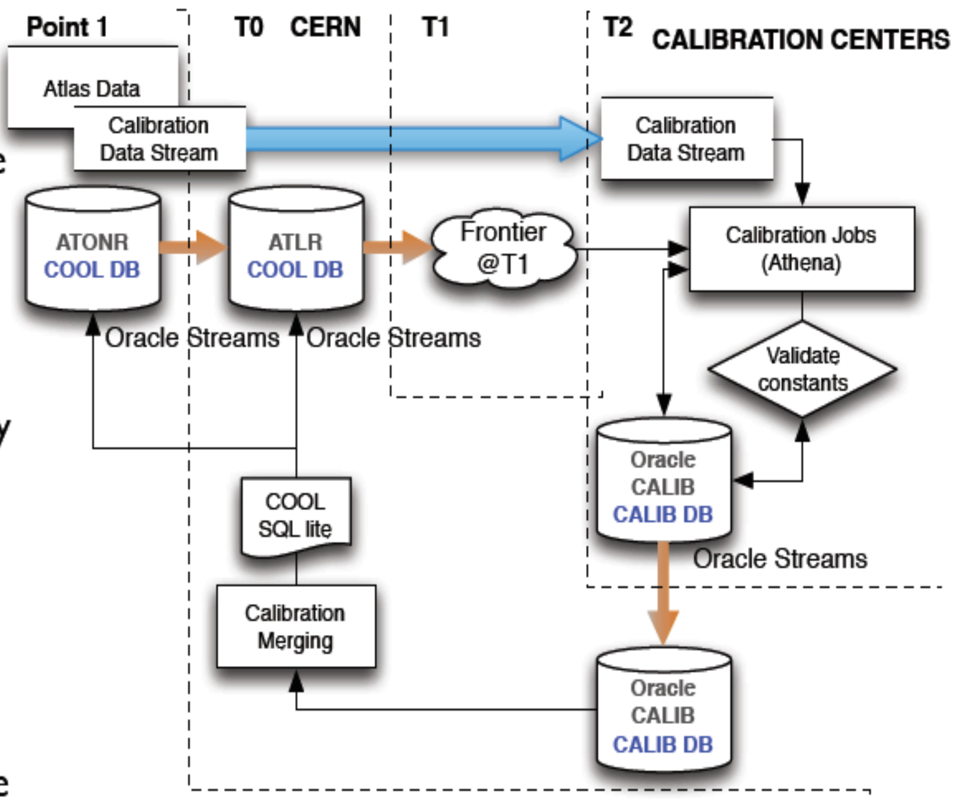
- We are in the process of assessing and revising requirements, real use cases and infrastructure
  - Focussing on the real uses (or requests) from the Data Preparation and Physics communities
  - Trying to simplify the system as much as possible
  - Reducing the cost in terms of Oracle infrastructure
- Proof of principle being worked out
  - Using Hadoop/HBase instead of Oracle back-end for data storage
  - Separating functions between files and database (structured storage)
- Difficult to assess manpower needs and availabilities in 2013 before the project is well defined
  - On the other hand availability of manpower influences project design

Tag Evolution Timescales

- Considering that Run2 should start at the beginning of 2015:
  - Jan-Jun 2013: tests of data formats, schemas, performance of upload, search and retrieve data on a reduced dataset (1 TB)
  - Jul-Dec 2013: implementation of the chosen solution on the (by then available?) CERN Hadoop cluster; adaptation or development of external services; upload of all existing data
  - Jan-Jun 2014: commissioning of the new system; performance optimization
  - Jul-Dec 2014: discontinuation of Oracle TagDB; commissioning with new cosmic-ray data
- This is a success-oriented schedule
  - It assumes that the Hadoop/HBase back-end is better than Oracle and scales to 100 TB by 2017 (end of Run2)
  - Also that there will be enough manpower and hardware resources for the new system

Muon Calibration Centres (1)

- Muon calibration processing runs in 4 different centers:
  - Rome, Michigan....and 2xMunich (RZG and LRZ which share 1 DB instance)
  - Every center hosts an instance of Oracle DB, where calibration constants are stored
  - Physics data are streamed towards calibration centers, every center handles a set of data and performs the reconstruction producing in output muon calibration constants like RT and T0s
- Oracle usage:
  - Oracle DB is used in every center, and data are replicated among different DB instances via Oracle Streaming software
  - CALIB@ATLR : contains 1 Schema / calibration center
  - @T0 a job "merges" all needed constants into an SQLite file (COOL format) which is then validated and inserted into production COOL servers (online and offline)



Muon Calibration Centres (2)

- Calibration centers do not profit of CERN external licenses for their Oracle installation
- Part of the data flow could be reviewed (Oracle streaming from calibration centers to CERN) if needed
  - Evaluate economical benefits of migrating to Oracle standard license
- LS1 is a good opportunity to investigate on possible (cheaper) alternatives
  - Usage of other free SQL databases
  - Usage of multi tier model for DB access (Frontier or CoralServer)
- A discussion on these topics is going on inside the Muon community

Today's conclusions

- Medium-term planning by the CERN IT-DB group
  - Not clear how much more Oracle will cost in the future – perhaps we should not be scared (yet)
- Deployment and access model for Geometry, Trigger and Conditions DBs
  - 4 Oracle+Frontier servers at IN2P3-CC, RAL, BNL, TRIUMF after mid-2013
  - DB Releases will be progressively phased out except for large-scale production and reprocessing campaigns – all remote DB access through Frontier
  - Conditions data files in CVMFS – drop HOTDISK
- Evolution of TagDB and its deployment model
  - Consolidation of data in Oracle at CERN+TRIUMF (real) and DESY (MC)
  - Proposal for an EventIndex based on Hadoop/HBase (2 years of work)
- Muon Calibration Centres at Tier-2s
  - Muon community will analyse the current model with 3 Oracle databases at Tier-2s and hopefully come up with a simpler scheme