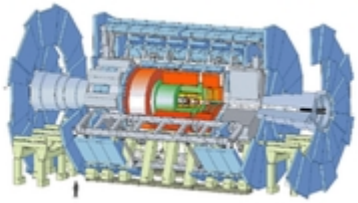
the ATLAS Experiment



Sliding window on the PVSS data on the ATONR database and ADCR application issues (plus other interesting DB issues)

Gancho Dimitrov (BNL)

Outline

- Developments / improvements since the previous SW week in Dec 2010

  - 12 months sliding window set to the PVSS (ProzessVisualisierungs und SteuerungsSystem) data on the ATONR database

  - new segment organization of the PanDA (Production ANd Distributed Analysis system) data

  - new procedures for better column statistics to reflect the reality.

  - The ATLAS geometry data to be replicated to the Oracle databases @ the T1 centers and configure FronTier to support it

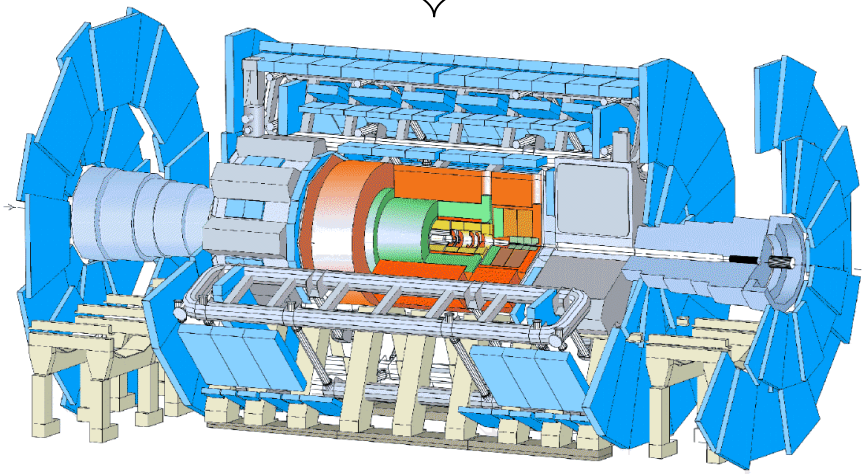
  - Conclusions

Introduction to the PVSS system and its use in ATLAS

PVSS (Prozessvisualisierungs und Steuerungssystem) is a control and data acquisition system being in use in the LHC experiments since year 2000.

PVSS by ETM

Thousands of data point elements



The ATLAS detector

PVSS Oracle archive -
keeps history of the detector status,
e.g. high voltages, temperatures

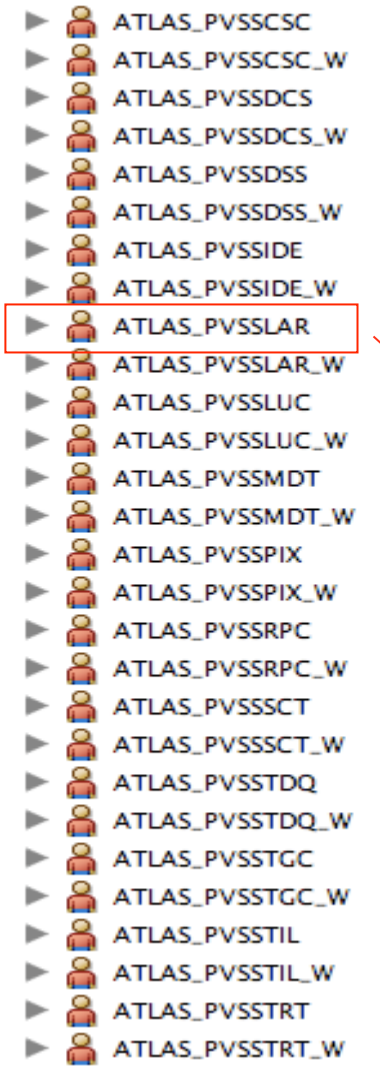
ATONR

ORACLE

The ATLAS ‘online’ Oracle DB

The ATLAS PVSS DB accounts and table desc.

- A database schema per subdetector (as total 14)



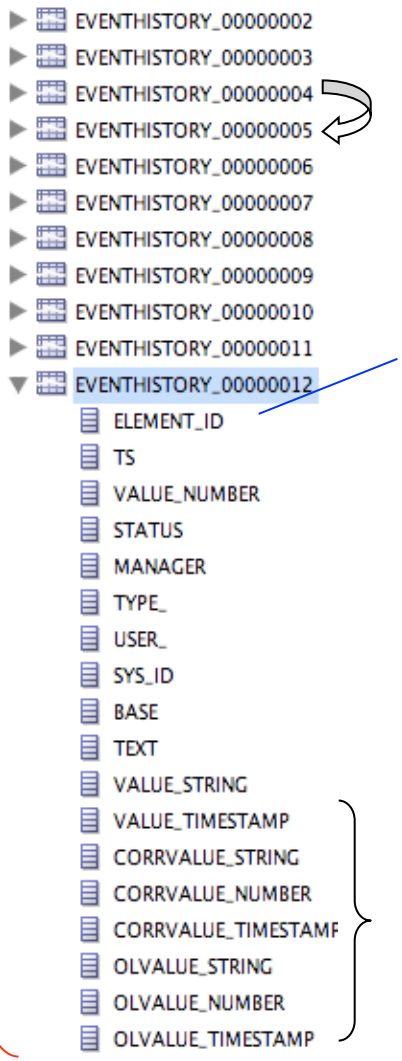


Table is 'switched' when it reaches a certain size and a view is updated to keep them together for the application to access the data (the EVENTHISTORY view)

Data point elements, in the LAR case are about 4500

Not used from ATLAS, get NULL values, thus do not take occupy space

The row length is in the range 55-60 bytes

Sliding window for the PVSS Archive on the ATONR

- An idea of keeping only the data of the most recent 12 months on the ATONR (sliding window) popped up naturally.

  The reasons are:

  - the operators in the ATLAS control room do NOT need to look furhter than 12 months in the past.

  - the complete archive is already on the ATLAS ‘offline’

  - the ‘online’ DB is vital for the datataking and is wise to be kept smaller in case of a need of recovery operation.

  The PVSS data (all tables and index segments) of 12 months occupy ~ 2.5 - 3 TB

Sliding window for the PVSS Archive on the ATONR (2)

1) That approach implies a move from the current « tablespace size threshold » to a « time interval » one

Successfully configured on 18th Jan 2011. On the 20th day of the month new tables apprear in their schemes. Depending on the average insert rate, for certain account this event will happen monthly, for others on few months and etc.

2) An important is to prevent table dropping on the source DB from being propagated on the destination DB.

A double protection is in place – a tagged session on the source DB and special code in the APPLY handler on the destination DB that discards any dropping table messages.

3) First major clean-up was finalized on 7th Feb (thanks to Marcin Blaszczyk)

The data from years 2007, 2008 and 2009 was removed from the ATONR database, leaving the ATONR with size of about 4TB

The PANDA ‘live’ and ‘archive’ data

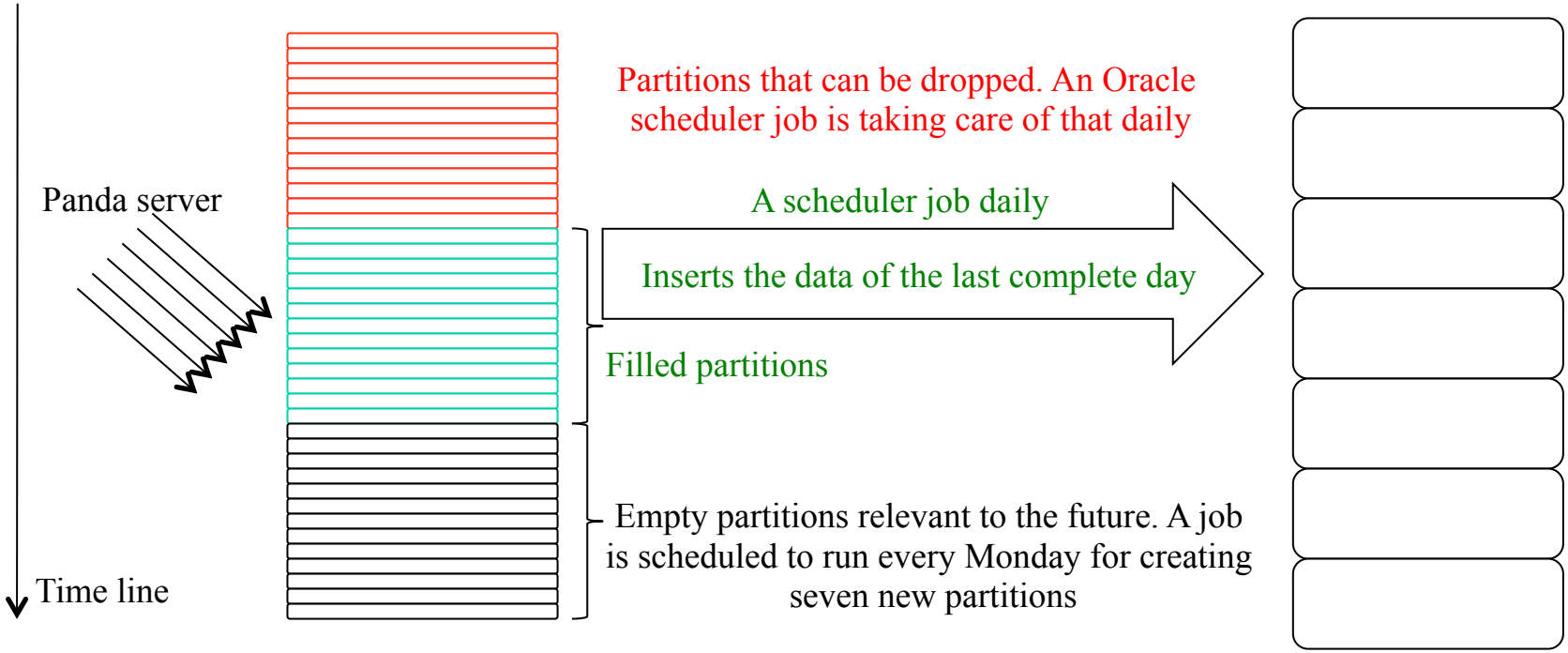
- The PanDA system is the ATLAS workload management system for the production and user analysis jobs
- All information relevant to a single job is stored in 4 basic tables. The tables are 4 because the most important stats are kept separately from the other space consuming attributes like job parameters, files details, inputs, outputs .. etc.
- The ‘live’ data is kept in a separate schema that keeps jobs of the most recent 3 days. Jobs that get status ‘finished’ or ‘failed’ are moved to the archive PANDA schema.

  ATLAS_PANDA => ATLAS_PANDAARCH

New data segments organization since Dec 2010 and further developed in Jan 2011

ATLAS_PANDA.JOBSARCHIVED4 table:
partitioned on a 'modificationtime' column.
Each partition covers a time range of a day
(the same to be applied for other tables as well)

ATLAS_PANDAARCH.JOBSARCHIVED:
range partitioned based on the
'modificationtime' column.



Important: For being sure that job information will be not erased without being copied in the PANDAARCH schema, a special verification PLSQL procedure of mine is taking place before the partition drop event!

Benefits after the changes

- High scalability: the Panda jobs copying and deletion is done on table partition level instead of row level as before
- Removing the already copied data is NOT an IO demanding (very little redo plus does NOT produce undo) as this is a simple Oracle operation over a table segment (alter table … drop partition …) and its relevant index segments
- Much better space utilization
- No need for indexes rebuild or coalesce operations for these partitioned tables.

A generic problem with the Oracle statistics gathering approach

- For queries that are interested in data of the most recent hours, often get non-optimal execution plan and thus consume a lot of resources.

  e.g. For the 'WHERE modiftime > SYSDATE - 1/2' the Optimizer considers that there are only few rows relevant to that condition even if the statistics are very recent (computed from the last night). In reality, for a ½ day in several different schemas we could get tens or hundreds of thousands rows. With the wrong statistics Oracle produces non-optimal execution plans.

  A real case is where more then two indexes exist and Oracle decides for the inappropriate one or when a join of two tables is needed, Oracle chooses NESTED LOOPs within a index range scan is taking place instead of HASH JOIN. That leads to much more buffer reads (respectively IO and CPU)

  Currently to stabilize the execution plan the queries are 'strength' with a lot of hints for instructing the Optimizer (e.g. INDEX_RS_ASC, NO_INDEX, CARDINALITY, USE_HASH ...etc)

A better fix would be …

- Explicit update of the timestamp column max value in the Oracle's data dictionary (as advised by the guru in Oracle performance tuning, Jonathan Lewis)

  - First beneficiary of that would be the PVSS Data Viewer tool (the next would be PanDA).

  - Currently tests are ongoing on the INTR database where a PLSQL procedure and a job of mine are changing the TS column max value of the latest PVSS EVENTHISTORY tables.

  - The test results are positive, thus, I plan all these components to be set on production ATLR database in the near future

The ATLAS geometry database

- Since 24th March the ATLAS geometry data is supported by FronTier / Squid at CERN (on later stage on the T1s as well)

- On 14th April the geometry DB schema has to be added to the production Oracle streams replication flow CERN => T1s

  - Certain modifications in the objects had to be done in order to be compatible with the Streams (e.g. elimination of the materialized views).

  Done successful tests on INTR = > INT8R replica.

Conclusions

- The ATONR database has a controlled max volume after introducing the 12 months sliding window for the PVSS data
- The ATLAS_PANDA schema is with increased scalability after setting up a set of Oracle jobs for adding, copying and dropping on partition level.
- Important work on adjusting the Oracle column statistics for the “data or timestamp” type columns to reflect the reality.
- Geometry database to be added to the production replica CERN => T1s, data to be served from FronTier as well

☺ A special “THANK YOU” slide ☺

A special, warm

OBRIGADO

and

*БЛАГОДАРЯ**

to

FLORBELA

for being a great colleague in all these years we worked together!

* Благодаря = ‘Thank you’ in my native alphabet and language (Bulgarian)