The ATLAS TAGS Database distribution and management

Operational challenges of a multi-terabyte distributed database system

VIEGAS, Florbela (CERN) MALON, David (Argonne National Laboratory) CRANSHAW, Jack (Argonne National Laboratory)
DIMITROV, Gancho (DESY) GALLAS, Elizabeth (University of Oxford) GAMBOA, Carlos (Brookhaven National Laboratory)
NAIRZ, Armin (CERN) GOOSENS, Luc (CERN) NOWAK, Marcin (Brookhaven National Laboratory) VINEK, Elisabeth (Universitaet Wien) WONG, Andrew (TRIUMF - Canada's National Laboratory for Particle and Nuclear Physics)

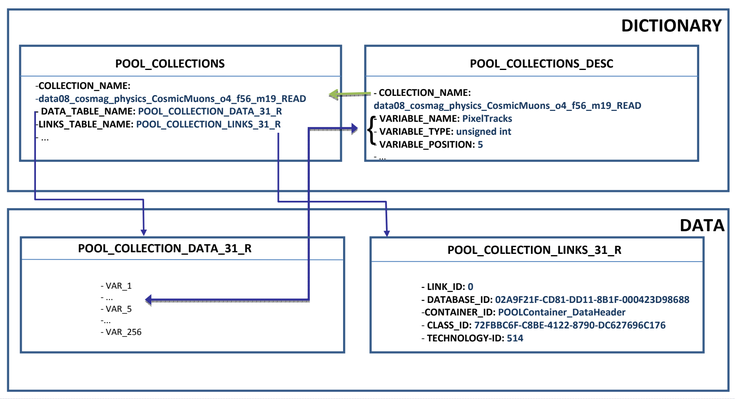
Computing in High Energy and Nuclear Physics
Prague | Czech Republic | 21 - 27 March 2009

1. The Challenge

The TAG files store summary event quantities that allow a quick selection of interesting events. This data will be produced at a nominal rate of **200 Hz**, and is uploaded into a relational database for access from websites and other tools. The estimated database volume is **6TB per year**, making it the largest application running on the ATLAS relational databases, at CERN and at other voluntary sites.The sheer volume and high rate of production makes this application a challenge to data and resource management, on many aspects. This poster will focus on the operational challenges of this system. These include:

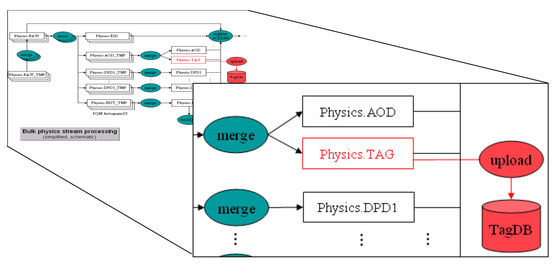
- uploading the data from TAG files to the Oracle database (at CERN and remote sites).
- distributing the TAG metadata that is essential to guide the user through event selection;
- controlling resource usage of the database,
  - from the user query load
  - to the strategy of cleaning and archiving of old TAG data.

2. Format of TAG data in the Database: the POOL Relational Collection Model



- Root files are read into POOL Relational Collection format (above figure).. Schema structure is very simple and flexible
- For efficient TAG upload and querying, a procedure call was added to make the Oracle table partitioned by run number, called POOL_COLLECTION_INIT
- For making the collections readable and scalable, the data tables are indexed completely on every column after upload.
- The collection name is taken from the dataset name, and two collections exist during upload: one for loading, and another for reading. Data is exchanged and indexed after run is completely loaded.

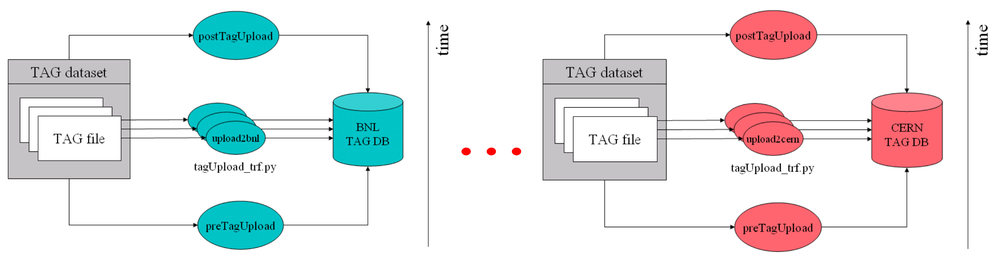
3.Upload and Distribution of TAG data using the T0 System (1)



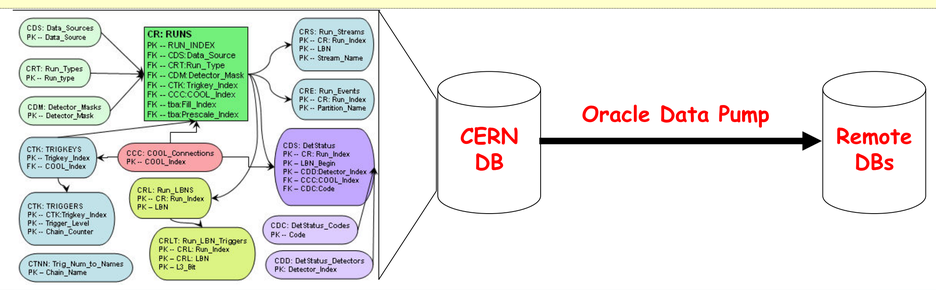
- TAGs are uploaded as they are produced
- T0 has defined tasks for each site to upload
- Connection is made directly to the database using POOL Collection utilities in an upload python transform script.

Distribution to Remote Sites

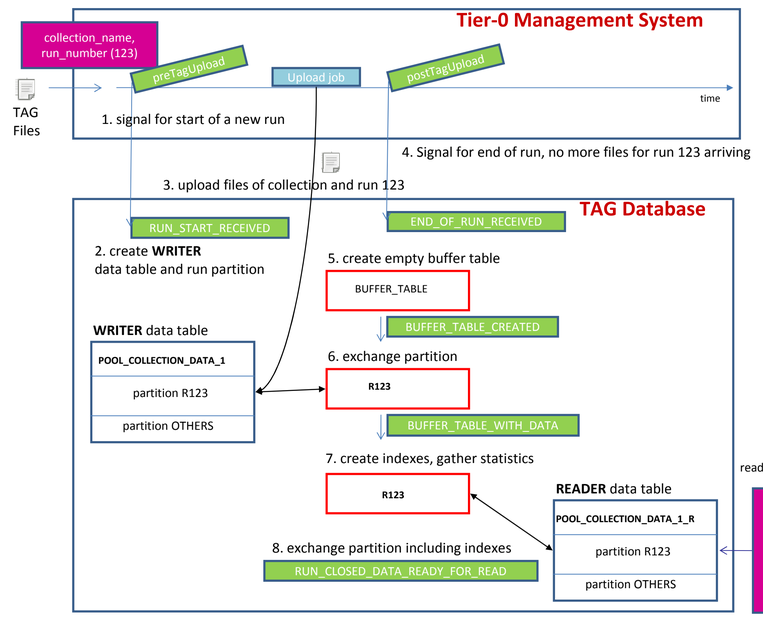
- Different tasks within T0 job system for each site and dataset, currently: CERN, TRIUMF,BNL and RAL (for testing only). DESY will be a future site.
- Description of preTagUpload and postTagUpload tasks in section 4.



- Run/LuminosityBlock Metadata (2) is exported regularly using Oracle Data Pump. This data is needed in the same database as the collections for the ELSSI website.

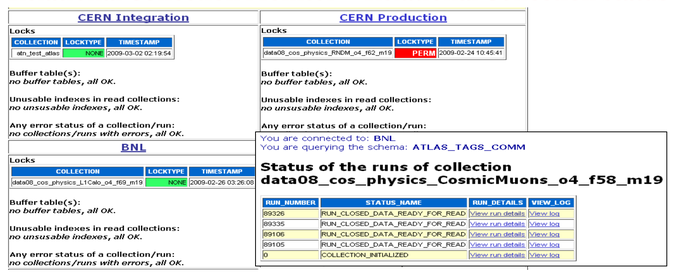


4. From TAG file to Relational Collection



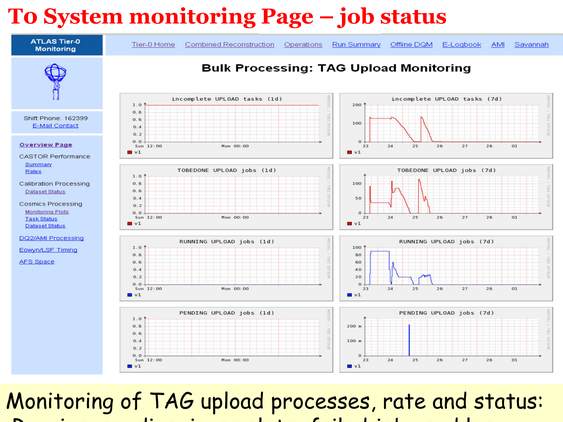
5.Monitoring TAG Operations

Relational Collection Monitor – Relational Collection status



•TAG upload dashboard: CERN, BNL, TRIUMF, RAL and in the near future DESY.
• Provides catalog and troubleshooting information on uploaded collections. Message logs of tasks are also accessible.

To System monitoring Page – job status

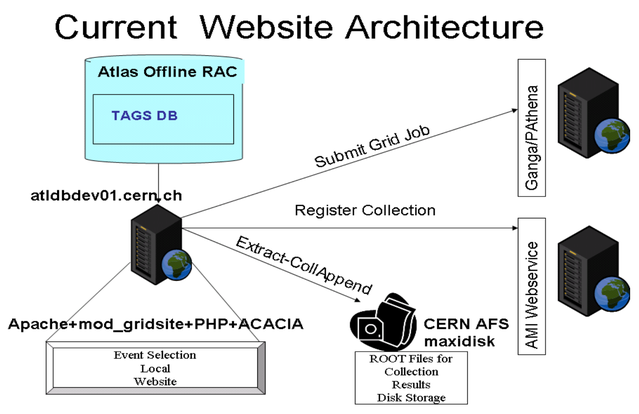


Monitoring of TAG upload processes, rate and status: Running, pending, incomplete, failed jobs and log output.

Monitoring the Oracle Databases – for performance and scalability

• Oracle Enterprise Manager is used for monitoring the behaviour of TAG databases.
• The local DBA for the remote sites will also actively monitor the database via OEM, Nagios and Ganglia.

6. Querying the TAG Database



- ELSSI : Event Level Selection Service Interface (3)
- **https://cern.ch/tagservices**

- Website developed for easy querying of the TAG database.
- Each site that hosts a database will setup this service locally
- Physicists are encouraged to use it as primary portal of access to the TAG database. Atlas VO Grid Certificate and Mozilla browser are needed for access.

Database Resource Usage Control

- Writing access to DB only from T0 processes, password is hidden, processes are throttled by the T0 system.
- Reading access mainly from ELSSI website, password is hidden, queries are controlled for resource usage.
- Other reading access from CollAppend uses database schemas with limited resources allocated.

Data Management and Archiving

- Collections are not updated and can be put in read-only mode once they are complete.
- Partitioning model by run number is crucial for identifying reprocessing passes and deciding on which data to keep.
- Data model and capacity planning foresee the storage of 2 reprocessing passes per year. Older passes will be purged from the database.

For more information, see CHEP 2009 presentations :
(1) The ATLAS Tier-0: Overview and Operational Experience by G.Negri (2) An Integrated Overview of Metadata in ATLAS by E.Gallas, (3) Event Selection Services in ATLAS by J.Cranshaw