Databases Overview

Andrea Formica, Shaun Roe,
Nurcan Ozturk & Elizabeth Gallas
for the Databases team

Database Operations and Conditions DB Developments during ATLAS Sites Jamboree and HPC Strategy Meeting
CERN, March 7th 2019

Outline

- Follow-up Review of Conditions for Run-3
- COOL-REST Prototype
- IOVDbSvc for REST Access to Conditions
- COOL Folder IOV Precision
- COOL Global Tags
- Frontier Analytics
- Frontier Operations
- Frontier Kibana Monitoring
- Squid SAM Tests
- Database Admins News
- EventIndex

Follow-up Review of Conditions for Run-3

ATDBOPS-26: Epic to cover the concrete recommendations of the 16 January 2019 review on database infrastructure for Run-3

Issues in Epic

ATDBOPS-27	Tune overlay tasks folder caching parameters to gauge effect on database	OPEN	Nurcan Ozturk
ATDBOPS-28	Investigate DB-intensive workflows apart from overlay	IN PROGRESS	Nurcan Ozturk
ATDBOPS-29	Explore the use of separate IOV/payload queries using COOLR	IN PROGRESS	Andrea Formica
ATDBOPS-30	Look at COOL-to-CREST conversion data size inflation	OPEN	Evgeny Alexandrov
ATDBOPS-31	Demonstrate IOVDbSvc reading of data from multiple data sources	OPEN	Shaun Roe
ATDBOPS-32	Investigate the possibility to throttle tasks based on DB overload alerts	IN PROGRESS	Nurcan Ozturk

Progress on Tickets

Nurcan Ozturk

- ATDBOPS-27: **Tune overlay tasks folder caching parameters to gauge effect on database**
  - 10min default cache length was overridden to 10s for some folders in overlay production in 2017 for the purpose of retrieving less Conditions data. Apart from DCS this would not work, resulted in redundant queries, bad caching. Need get this setup removed from the jobOption, Andrea plans to offer a list of folders where the cache length can be overridden. Will request this removal and just use the default for now.
- ATDBOPS-28: **Investigate DB-intensive workflows apart from overlay**
  - Discussed with Richard H. and Peter O. on the sparse DRAW reco, all the tasks I collected in the spreadsheet cover already such tasks; DRAW_TAUMUH (the most sparse one, no alerts), DRAW_RPVLL (a good sparse one, with alerts), DRAW_ZMUMU (with alerts). Work in progress to study these tasks.
- ATDBOPS-32: **Investigate the possibility to throttle tasks based on DB overload alerts**
  - In progress; a new global share will be defined and the tasks appear in the Frontier alerts will be directed to this global share thus throttled by the share instead of the current way by the cap on the number of running jobs.

COOL-REST Prototype

Andrea Formica

Address the caching problem

- Several proposal from Jan review follow up : separation of iovs / payload queries
- How to implement this with minimal changes ?
  - After discussions with COOL experts we decided to go for a simple first attempt
  - Deliver to client (Athena IOVDbSvc) a list of IOV ranges
  - PL/SQL query implemented
  - COOLR endpoint implemented
  - Further tests are needed
  - Select the folders for which is most useful to apply this procedure
  - Some problems detected for usage of nanosecond precision (when not needed)

Browsing COOL data via COOLR

- Progress in the WEB-UI that could replace cool tag browser in the Run3 perspective
- Development of new simple python scripts to spot differences in tag content

IOVDbSvc for REST Access to Conditions

Shaun Roe
Evgeny Alexandrov
Mikhail Mineev

IOVDbSvc

- Refactoring continues
- Extensive unit tests introduced in 21.9 and master, to compare
- A simple job accessing the CREST database has been demonstrated to work
  - **Without** changing client code
  - Only a single IOV so far
- Aim to run q-tests from the CREST database March/April

COOL-CREST conversion

- Prototype tool exists, analogous to AtlCoolCopy
- Tests on real data (e.g. a convert all data necessary for a q test) planned in March

COOL Folder: IOV nsec precision !?!

Elizabeth Gallas,
Andrea Formica,
Gancho Dimitrov

- COOL DB : allows ‘nanosecond’ precision for IOV ranges (in time-based folders)
  - This is important for some conditions information … i.e EventVeto, LB to time conversion, etc.
  - But such precision is unnecessary for other conditions … i.e. HV, LV !
    - The ‘time’ used is probably the insertion time of the data
      rather than a precise time associate with the inserted condition.
    - When there are lots of channels all with slightly different IOVs
      the number of unique IOV ranges of the folder explodes:
      In one folder, we found 2700 IOV ranges becomes 350k ranges !
- CREST resolves this in a natural way: all channels aggregated into JSON object with a common IOV
  → **the challenge is converting existing COOL data with such irregular IOVs**
- What is the extent of the issue ? i.e. How many folders are affected?
  - CONDBR2: currently has 374 ‘time’ based folders
    - Found 226 folders use nsec precision in IOVs (link)
    - Count: ‘nsec’ precision folders per subsystem →
      out of all time-based folders (it's not all time folders)
  - → need to address folder by folder
    (for folders which need to be converted …)

TABLE_OWNER	NSEC_FOLDERS	TIME_FOLDERS
ATLAS_COOLOFL_DCS	11	58
ATLAS_COOLOFL_LAR	1	1
ATLAS_COOLOFL_MUONALIGN	1	8
ATLAS_COOLOFL_TRIGGER	2	2
ATLAS_COOLONL_FWD	21	21
ATLAS_COOLONL_LAR	35	67
ATLAS_COOLONL_SCT	16	22
ATLAS_COOLONL_TDAQ	41	43
ATLAS_COOLONL_TRIGGER	98	106

COOL Global Tags

- Global Data Tags
  - CONDBR2-BLKPA-2018-12: used from 24-Sept to end of Run 2
  - Created or In preparation:
    - CONDBR2-BLKPA-2018-13 (use for Pixel paper and possible partial reprocessing) → "Current"
    - CONDBR2-BLKPA-2018-14 (BK for HI 2015,2016 reprocessing) → "Next"
    - CONDBR2-BLKPA-2018-15 (BK for Run 2 -- updates from various subsystems)
- Global MC Tags
  - Created: New IOVs for various 2018 beam configurations
  - Preparing: New MC in preparation for HI production
  - Collecting: Updates to conditions from many subsystems
- Operations: Validation and Tag Collection (*problems occurred recently with some "new" tags*)
  - Plan: Move from collecting updates from subsystems to a JIRA based system
    Need to include some 'sign-off' mechanism (like DQ)
- AFS phaseout: Many tools and operations have AFS dependence
  - Goal: transition by the end of 2019 !?!
    - Has challenges and uncertainty: can EOS handle Tier-0 parallel access? etc.

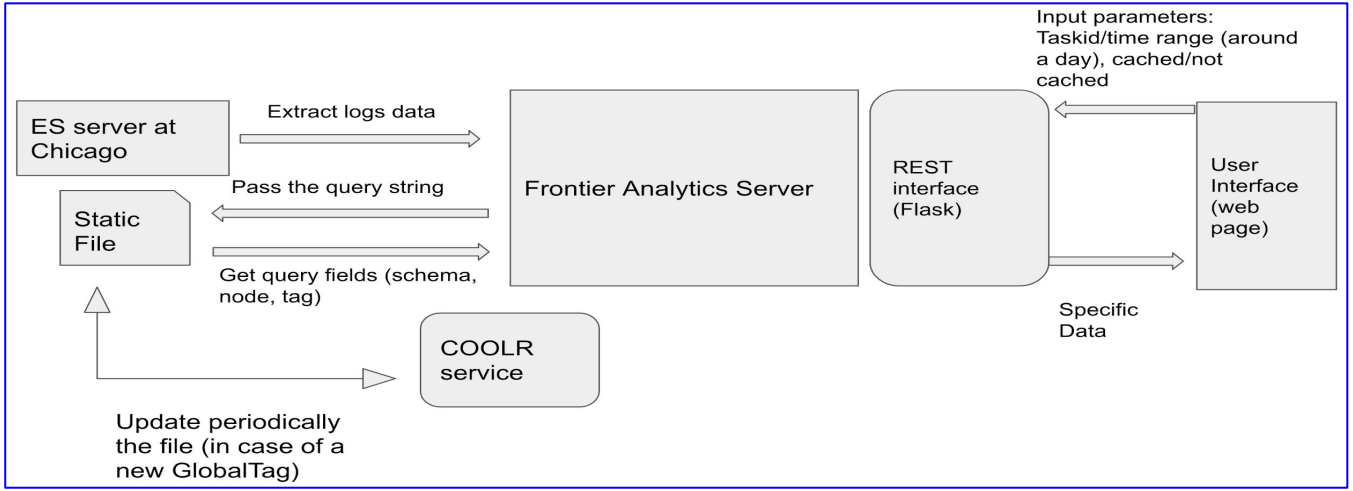
Frontier Analytics

Millissa Si Amer, Andrea Formica, Nurcan Ozturk

- Motivation was to understand the caching efficiency of Frontier-squid system by monitoring different tasks running in parallel regularly.
- A JupyterNotebook framework in Chicago was set up and used for studying and comparing different tasks.

Overlay task / reprocessing task (DRAW_RPVLL)

Conditions Folder	# queries	Different queries	Different payloads
/sct/dag/config/chip	5K / 358	1K / 237	3 / 2
/mdt/t0blob	8.5K / 44	7.5K / 1	1 / 1
/pixel/pixacalib	8.5K / 0	7.6K / 0	1 / 0



Millissa is now developing a real-time Frontier analytics system to study all the tasks running in parallel for a given Frontier(s) and monitor the caching efficiency per folder.

Frontier Operations

Alessandro De Salvo

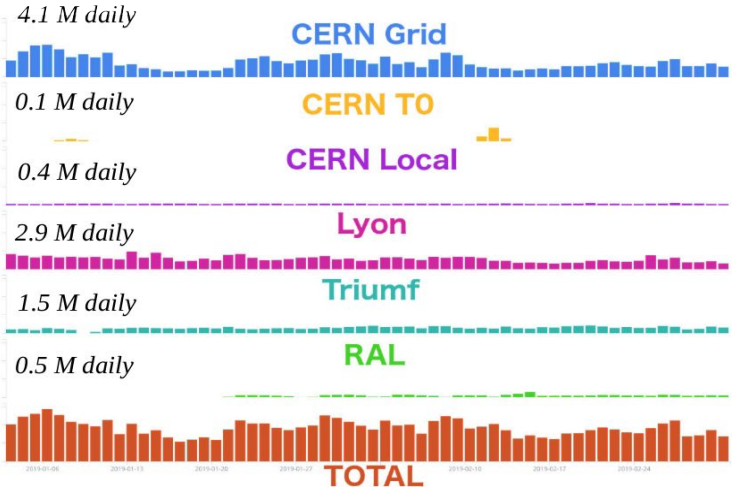
- Site status updates (since Dec 2018)
  - CERN
    - Squid-4 deployed on all launchpads, no problem observed so far
    - Problems with CI jobs
      - The problem was tracked down to be related to the squids running at Wigner with jobs in Meyrin
      - Temporary solution: reconfigure the CI jobs to use the Meyrin squids only
      - Long term solution: use proxy autoconfiguration
  - Lyon
    - Isolated incident: spontaneous I/O on backend storage system (SAN) created long responses to tomcat servlets and rejection of queries, now fixed
  - RAL
    - 3 new services running on the new infrastructure already in place
      - One service already in production, the other 2 will be in production soon, replacing the old ones
- Proxy autoconfig
  - The goal is to avoid cases when jobs in Wigner uses Meyrin squids and vice versa: automatic squid selection based on the client location
    - Implies a change in AGIS (has_proxyconfig) and changes in autosetup/ALRB
    - AGIS flag ready by last week, autosetup/ALRB implementation will start next week
- Backup Proxies
  - In production since a few months
  - Off-site squids cleanup in AGIS completed
  - The next step is to inhibit the direct connections to launchpads from nodes

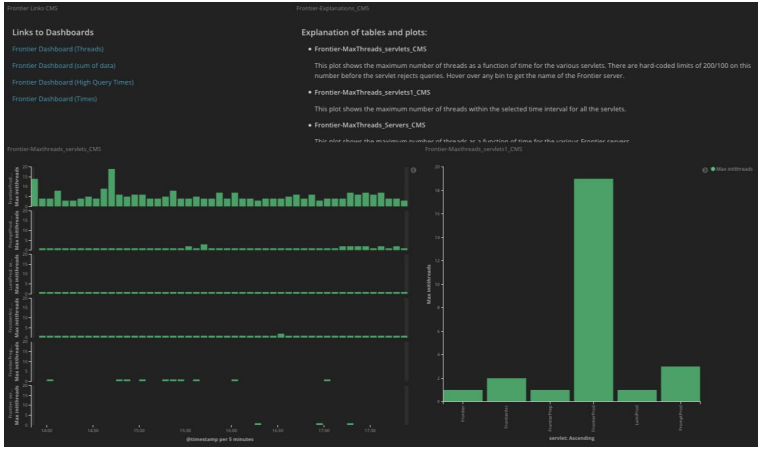
Frontier Kibana Monitoring

Julio Lozano Bahilo

- **Statistics:**
  - 9,3 M queries processed daily
  - 0.19% of queries failed on a server
- **Index:**
  - Switched to frontier (was frontier-new)
- **RAL servers:**
  - lcgvo-frontier05 monitored since 21st Jan
  - lcgvo-frontier06 and lcgvo-frontier07 soon
  - Dashboard already includes RAL site at the same level

- **CMS servers:**
  - 7 servers, though 3 are for testing purposes
  - All of them currently monitored
  - About 4M queries handled daily
  - New 'Frontier Dashboard CMS'
    - Basically a replica of the ATLAS Dashboard
    - Emphasis on servlets





Squid SAM Tests

Michal Svatos

- Tests were stopped for couple of years and started again few weeks ago
- They are testing access to launchpads via squids
- Monitoring link
- What each SAM status usually means:
  - CRITICAL: "NO PANDA RESOURCE FOUND"
  - WARNING: failed to contact one of the launchpads via one of squids
  - OK
- Dealing with the errors (a question similar to failovers):
  - What kind of expertise is needed to help sites fix the failures?
  - Who and how will follow up (depending on the expertise needed)?

Database Admins News

Gancho Dimitrov, Andrei Dumitru

- **CERN IT-DB database support**: best effort during LS2.
- **Oracle 18c version:** validate applications on testbed DB and upgrade production DBs by Q3 2019.
- **Oracle GoldenGate data replication:** ongoing discussions about its future in ATLAS
- **EventIndex Oracle setup**: has got 170B event records (6TB disk space). Keeps on working efficiently.
- **Event WhiteBoard:** new database schema agreed with the Event Streaming Service developer Wen Guan. DB objects inherit some of the interesting solutions of the previous schema.
- **Rucio**: sanitised DB schema on the prod DB. Characteristics presented on the second Rucio community workshop (Oslo, Norway)
- **New Small Wheel:** along with the NSW validation and assemble, its DB applications become more and more important. NSW documents are stored into the database, currently 36K (80% with size smaller than 100KB).

EventIndex

- EventIndex workshop was on Tuesday morning
- Operation status: generally smooth in the last few months
  - Usual (occasional) failures in data access due to site problems or file corruption:
    Identify → report to site or DDM → fix site or recover data → iterate until success
  - A few special cases:
    - New request to index MC HITS for Event Service datasets: check for duplicates
    - Indexed recent data12 Bphysics reprocessing : Run 1 indexing still working OK!
- New developments in HBase/Phoenix (for Run 3):
  - Tests of ingestion and search rates provide reasonable performance
  - Detailed schemas now worked out
  - Need to implement all (old and new) use cases and test again
  - Work starting on clients
  - Target for start of parallel operation with current system: end of 2019 (no need to rush)
- Next (long) EventIndex workshop on 3-5 June at LAL Orsay
- Feedback or special indexing requests welcome: JIRA ATEVTINDEX or email