LESSONS LEARNED

E. Orru' ASTRON

https://miro.com/app/board/uXjVLexIDc4=/

Quick look to the LTA archive of LOFAR

- 3 long term archive sites > ingest and downloads simultaneous
- Archive growth 5 PB/yr ~ 60 PB.
- Data products uncalibrated data and PSR fits
- State of the art algorithms & pipelines used by individuals
- User access via internet

Quick look to the LTA ambitions LOFAR 2.0

3 long term archive sites – **possibly more will be added?** > ingest and downloads simultaneous

Archive growth > LOFAR

Temporary Data products (intermediate calibration products)

Final Data products long term (science ready products)

User access via internet? Or what else?

State of the art algorithms & pipelines available in production environment

Outline

Point of view:

- Users
- Operations
- Developers
- Reporting

Brainstorm with you about mitigation actions and possible solutions.

Pov. User

STAGING

Slow & unreliable staging

Since upgrade at all sites overall performance improved

Communication in fixing problem not optimal

Lack Monitoring tool: performance, staing queue and status

Responsiveness at some site has been very effective.

PROCESSING

processed about 850, 2700 and 3400 datasets at poznan, sara and juelich respectively

LOFAR2 archiving data twice as fast as LOFAR1. LTA to be able to stage data faster than they ingest it so that people can keep up with the datarate.

Pov. Operations

INGEST

- LTA different infrastructure impact operations
- Low-level processed data big size - slow download & costly to save long term
- Control on Ingest queue developed over years
- Lack Monitoring tool: performance, staing queue and status

STAGING/PROCESSING

- Discovery and data access not user friendly
- No offline processing at LTA in production system (possible LDV one site)
- Data Model not optimal for discovery
- LOFAR2 archiving data twice as fast as LOFAR1. LTA to be able to stage data faster than they ingest it so that people can keep up with the datarate.

Pov. Developer

KNOWLEDGE

LTA originally developed and maintained by Astrowise and then continued by an Astron Sysadmin

knowledge and background information gone

especially about the code structure, the datamodel , procedure when the datmodel is updated etc.

SCATTERED RESPONSIBILTES
- Astron sysadmin - LTA
- Astrowise: maintenance Oracle DB & basic support
- Team Rainbow: Update(s) of the LTA software
- SDCO: Update main HTML page

MAINTENANCE/SUPPORT/DATABASE

The LTA software is stored in gitlab but there is no CI/CD. Manual upgrade of software is required.

Updating the datamodel via updating the datamodel in - to run manual update by running generated SQL scripts.
Deployment very risky

User and project management via old IDM might be headache now.

Oracle DB knowledge within Astron?

Not enough testcases available. Current integration tester to start manually but there is a lot of code which is/look obsolete. Due to lack of knowledge to 'dangerous' to clean this up.

Pov. Developer

LTA DATA.

Early days LOFAR, dataproducts were "thrown" into the LTA – high overhead when reprocessing and using/updating the current metadata into the LTA (duplicate dataproducts, SASids which are Unspecified or Unspecified and Pipeline or are associated with wrong projects).

SIP files cannot extract from the LTA itself.

The ingest of data goes via a SIP file. Creating SIP file quite cumbersome for clients. What if we would have the metadata described in the dataproducts itself?

CODE

The current python code does not have an easy (eg REST interface) to query or do updates of the dataproducts.

Removing of dataproducts was done by "buggy" script (by a software developer). Now functionality LTA-toolbox

Pov. Reporting to stakeholders

Request statistics for access LOFAR data products: LTA and VO => authentication vs non authentication.
Need of filtering and extra processing to isolate the meaningful information.

Measuring access and mapping that to individual users is very challenging.

Data products are provided through several access points, agreements (GDPR compliant) need to be made with third parties for sharing user access statistics.

Not automatising the metrics for usage statistic leads to high overhead for reporting

Conclusion:
Let’s write them together looking at the future challenges