Towards Dynamic Catalogues

Bart Scheers[1,2], Fabian Groffen[2], and the TKP Team[1,3]

1. Astronomical Institute "Anton Pannekoek", University of Amsterdam, NL
2. Cenrtum Wiskunde & Informatica, Amsterdam, NL
3. ASTRON, Dwingeloo, NL

LOFAR,

the next-generation radio telescope, is sensitive in the low-frequency regime of 30 – 240 MHz and designed to carry out unique science:

high-speed all-sky surveying
searching for rapid transient and variable sources
cataloguing the millions of sources and their millions of measurements

As a consequence LOFAR is going to produce tens of terabytes per day. High-cadence data rates of tens of gigabits per second are neither exceptional. Storing these huge volumes of scientific data requires unique database management systems that are, moreover, able to query the data scientifically with acceptable response times.
This poster shows how the Transients Key Science Project **[1]** of LOFAR approaches these challenges by using column-stores, sharded databases and the new array query language SciQL (pronounced as 'cycle').

Source Association Parameters and Variability Indices

determine whether sources may be genuinely associated, and if so, whether they show variability in their light curves, respectively. The values of the parameters and indices are related to the corresponding probabilities of the hypothesised property.
Here, one of the association parameters, the dimensionless distance, is shown. The positional difference of a pair of sources weighted by their errors resembles a Rayleigh distribution.

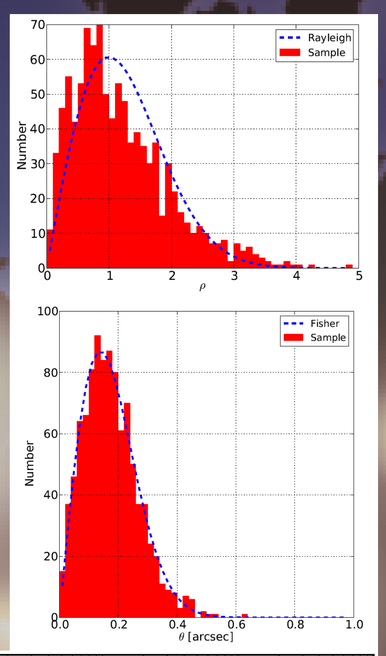
$$r_i = \sqrt{\frac{(\alpha_i \cos\delta_i - \alpha^\star \cos\delta^\star)^2}{\sigma_{\alpha_i}^2 + \sigma_{\alpha^\star}^2} + \frac{(\delta_i - \delta^\star)^2}{\sigma_{\delta_i}^2 + \sigma_{\delta^\star}^2}}$$

The *magnitude* of the flux variability of a source can be expressed as the ratio of the sample flux standard deviation, and the sample arithmetic mean flux, given by

$$V_\nu = \frac{1}{\overline{I_\nu}} \sqrt{\frac{N}{N-1}(\overline{I_\nu^2} - \overline{I_\nu}^2)}$$

The second indicator, which expresses the *significance* of the flux variability, is based on reduced χ^2 statistics. We assume that the weighted average flux value is a fitted parameter, so that the number of degrees of freedom is $N - 1$. It is given by the sum of the squared deviations from the weighted average weighted by the errors, and divided by the number of degrees of freedom flux, given by

$$\eta_\nu = \frac{N}{N-1}\left(\overline{w_\nu I_\nu^2} - \frac{\overline{w_\nu I_\nu}^2}{\overline{w_\nu}}\right)$$



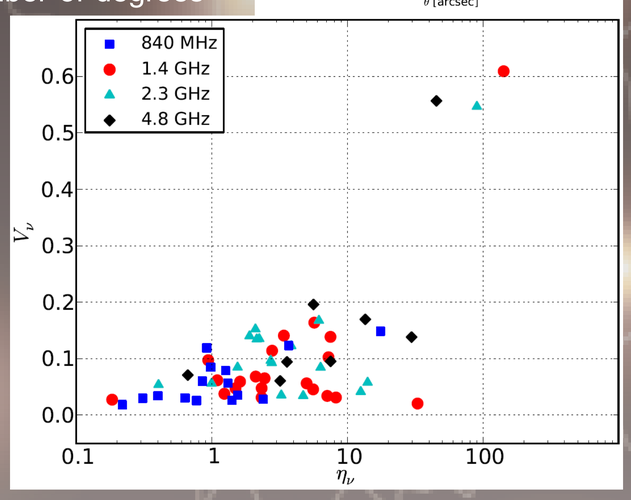


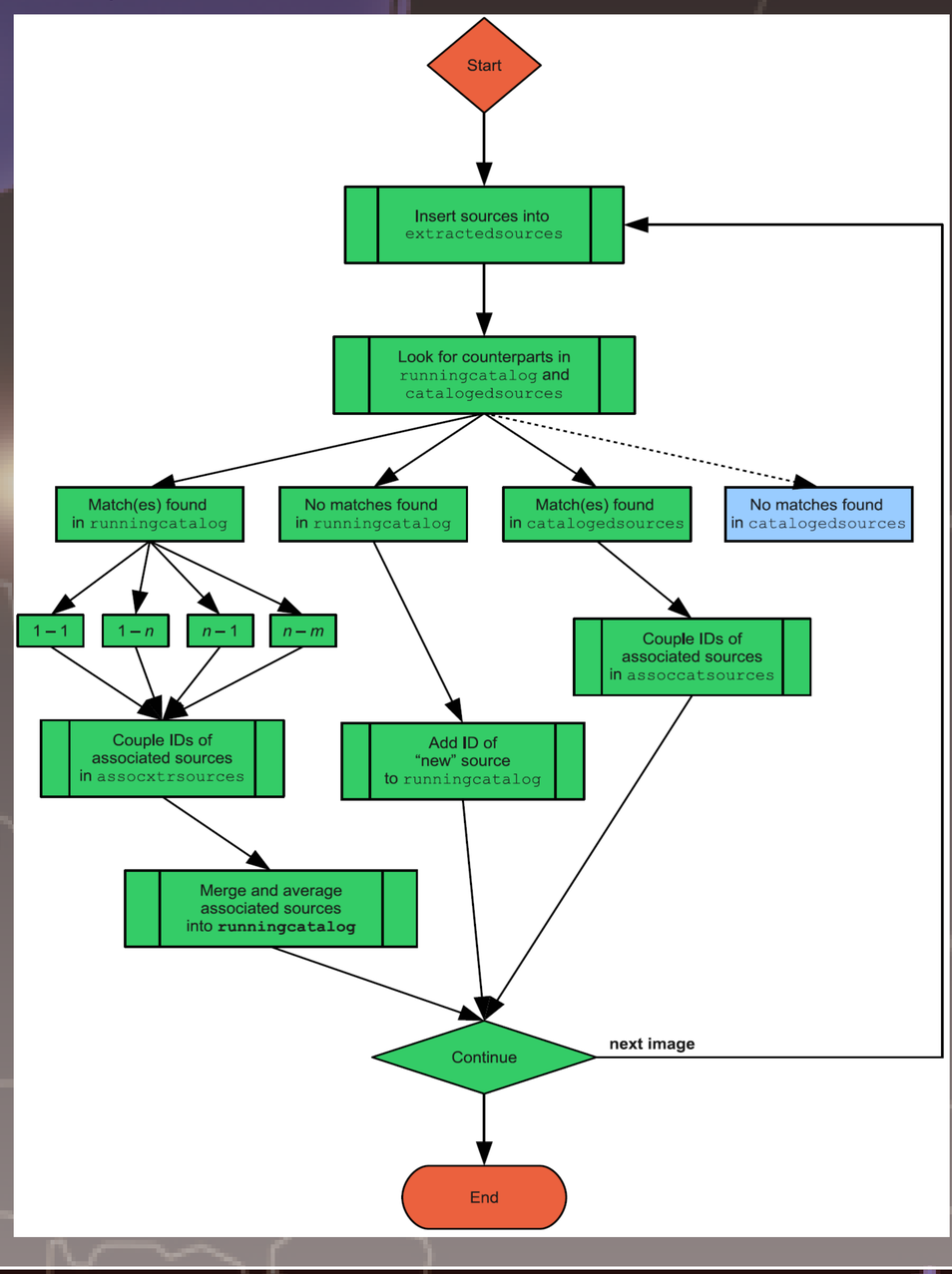
Fig. 2. A sample of 1000 images, each containing 64 sources, was processed in the Transients Pipeline for source association. **Upper**: The dimensionless distance of associated sources follows a Rayleigh distribution. **Middle**: the distance (arcsec) of the counterparts shows the Fisherian distribution. The **bottom** panel shows the variability indices for the sources in the field of view of GRB030329 different frequency bands of WSRT observations. A single symbol corresponds to a source and all its associations. The number of times a source could be associated is not displayed here. The afterglow of GRB030329 reveals itself by the three data points in the upper right region.

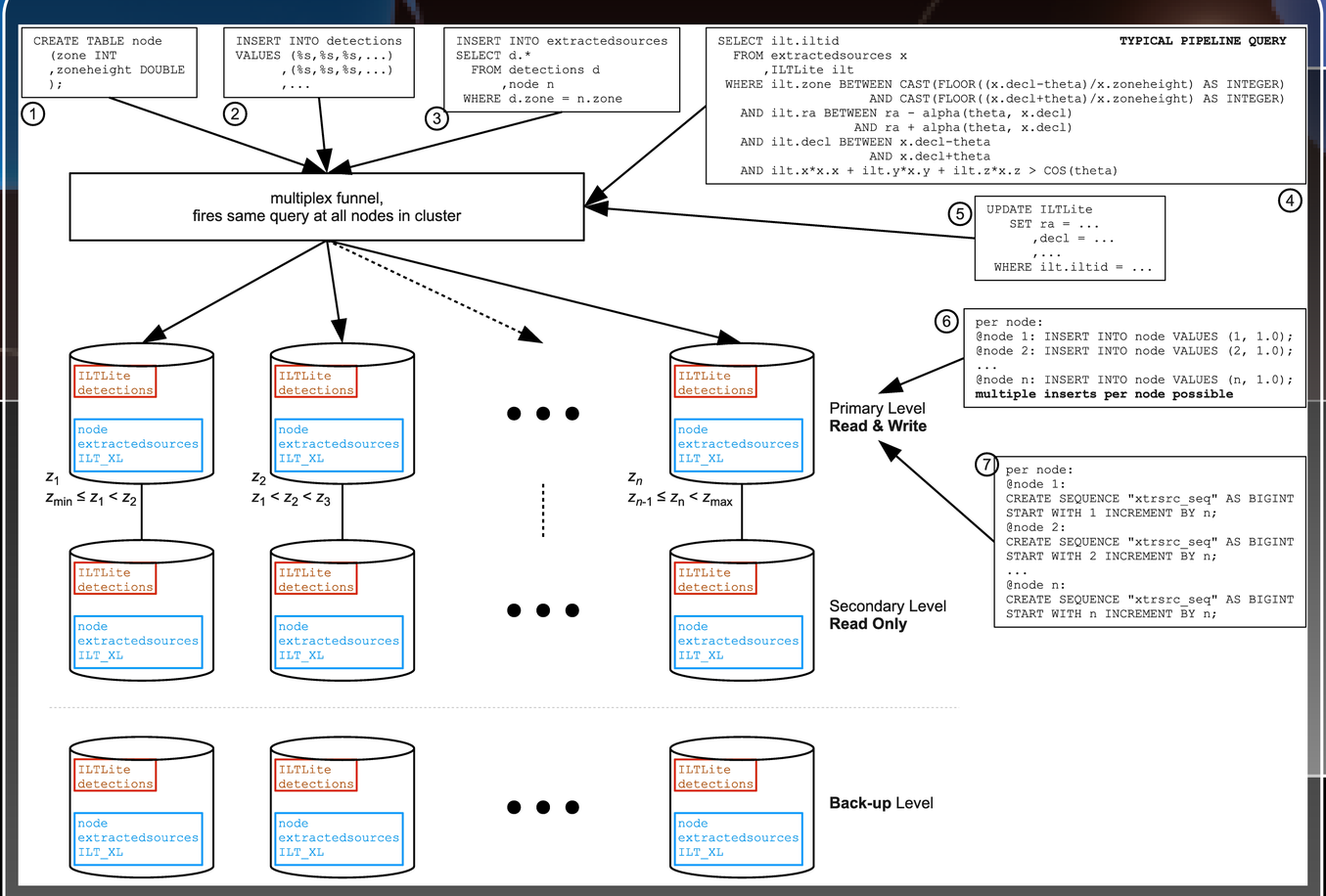
MonetDB, the open source column-store,

is fundamentally different from design than the classical row-store relational database management systems (RDBMSs), like MySQL or Postgres, but all are interfaced with the same Structured Query Language (SQL).
Direct consequences are that queries only touch the relevant columns, and when in contiguous memory it allows compression and good cache-hit ratios. Furthermore, MonetDB's kernel is a programmable relational algebra machine operating on "array"-like structures, exactly what CPUs are good at.
The experimental SciLens platform is a 330 node, 4-tier locally distributed infrastructure based on MonetDB technology and focusses on massive I/O, instead of raw computing power. The SciLens infrastructure is envisaged to be the prime choice for a scalable LOFAR light-curve database **[2]**.

The Transients Database

forms the heart of the Transients Pipeline, the software framework that aims for detecting transient and variable sources in the high-cadence calibrated LOFAR images.
During LOFAR observations, the Transients Pipeline stores *all* sources extracted from the images in the Transients Database. While observing sources are cross-matched with counterpart candidates in the spatial, spectral, polarisation and temporal domains of LOFAR and non-LOFAR catalogues as well. Crucially, to keep this query processing relatively constant over time, cross-matching is done against a statistical representation of the LOFAR catalogue (the so-called *running catalogue*), instead of the whole data volume (available in *extracted sources*), see Fig. 1 and **[4]**.
The *running catalogue,* a statistical summarisation of the millions of sources and their millions of measurements, serves as a global all-sky model as well, that is being used in the image calibration steps. The catalogue and the sky model will evolve and improve over time.





A Sharded Database for LOFAR Light curves

keeps the entire volume quickly accessable without replication as opposed to a distributed database environmentwill restrict the volume of replicated data. Queries are fired at all machines in the cluster through the multiplex funnel. In this set-up data is sharded by declination zones. Joins with steering tables, on all nodes, determine at which node a query is actually executed.
The tables in red boxes cover the whole sky, whereas the tables in the blue boxes cover only parts of the sky which are bound by the declination zones.
ILTLite is the "stripped" LOFAR Catalogue and contains all the unique sources (~10^8), i.e. the positions where at least once a LOFAR detection was made, but nothing more than that. This table is fully replicated over all the nodes, and it has only the essential columns needed for source association, in order to fit in memory. The ITL_XL tables contain all the unique sources within the given declination zone, and has all the columns. By joining the (pipeline or user) queries with the node table on zone, we do not have to replicate more data than the stripped ITL Catalogue.
SciQL [5, 6 & 7] eases the scientifically very relevant light-curve analysis by query window processing, where moving averages, Fourier Transformation, correlation and convolution are directly applied inside the database engine.

References

[1] Fender, Wijers, Stappers et al. (2007), PoS, p.30; **[2]** **www.scilens.org**; **[3]** Swinbank, 2010, ISKAF2010, 82; **[4]** Scheers, (2011), PhD Thesis, UvA; **[5]** Kersten et al., 2011, AD2011, 12; [6] Zhang et al., 2011, IDEAS2011, 10; **[7]** See Zhang's O29 talk at ADASS XXI