Towards Dynamic Ligh-Curve Databases

Bart Scheers

Centrum Wiskunde & Informatica, Amsterdam
Astronomical Institute "Anton Pannekoek", University of Amsterdam

May 6th, 2013

Outline & Future Key Developments

- LOFAR Status and future development
- Move key-science frameworks more upstream in pipelines
- Distributed Databases, use intelligence and autonomy of storage devices
- Array based query processing
  - Enhances data mining of PB dynamic catalogues
  - Fits to multi-dimensional datasets, e.g. MS, HDF5, FITS
  - Alleviates reprocessing and reloading of "raw" data
- Simultaneous multi-messenger observations
  - If catching all is not possible, only pick the low-hanging fruit

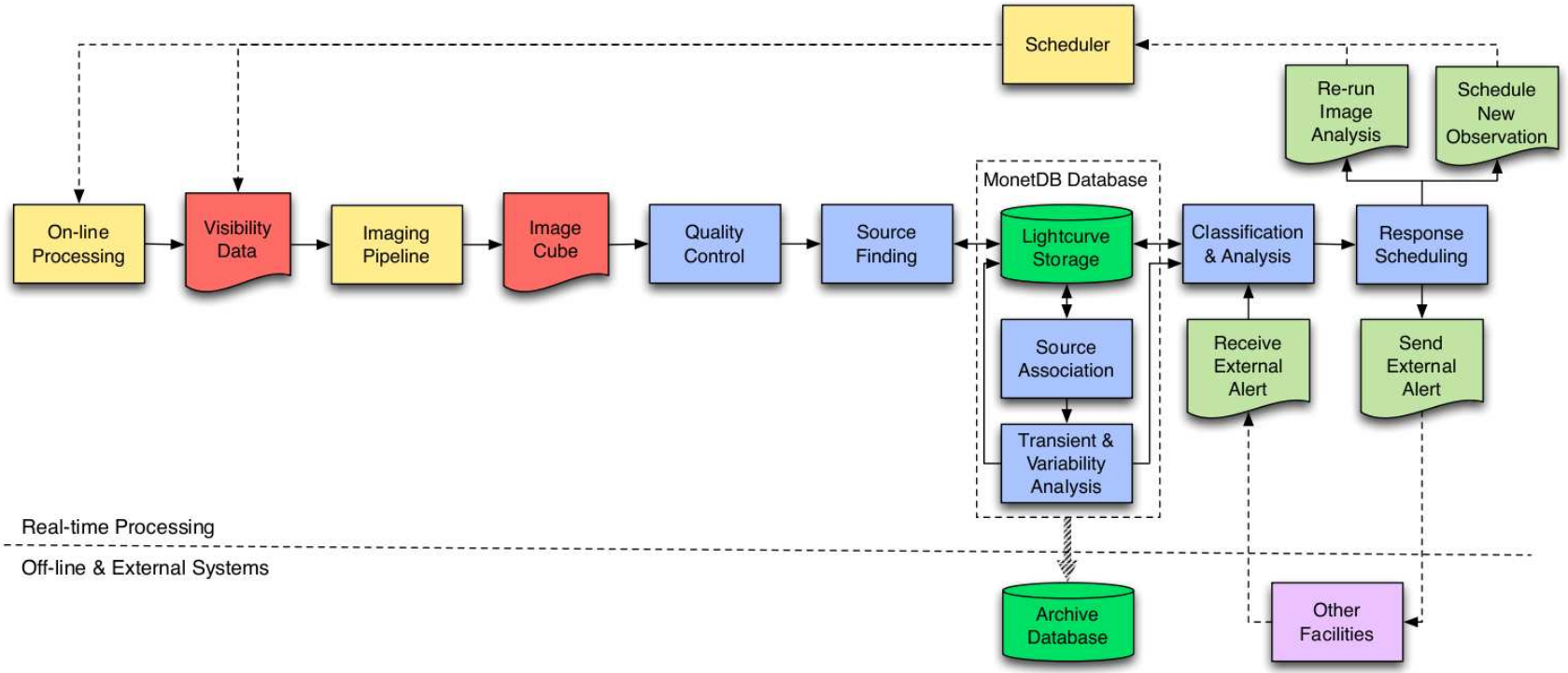
LOFAR Status – Characteristics

- Raw data $\sim 25\,\mathrm{TB/hr}$
- Distinct sources: $\sim 10^7 - 10^8$,
  - which are revisted many, many, many times
- Source properties reduce to $50 - 100\,\mathrm{TB/yr}$
- Peaks over 10,000 sources per second

LOFAR Status – Characteristics

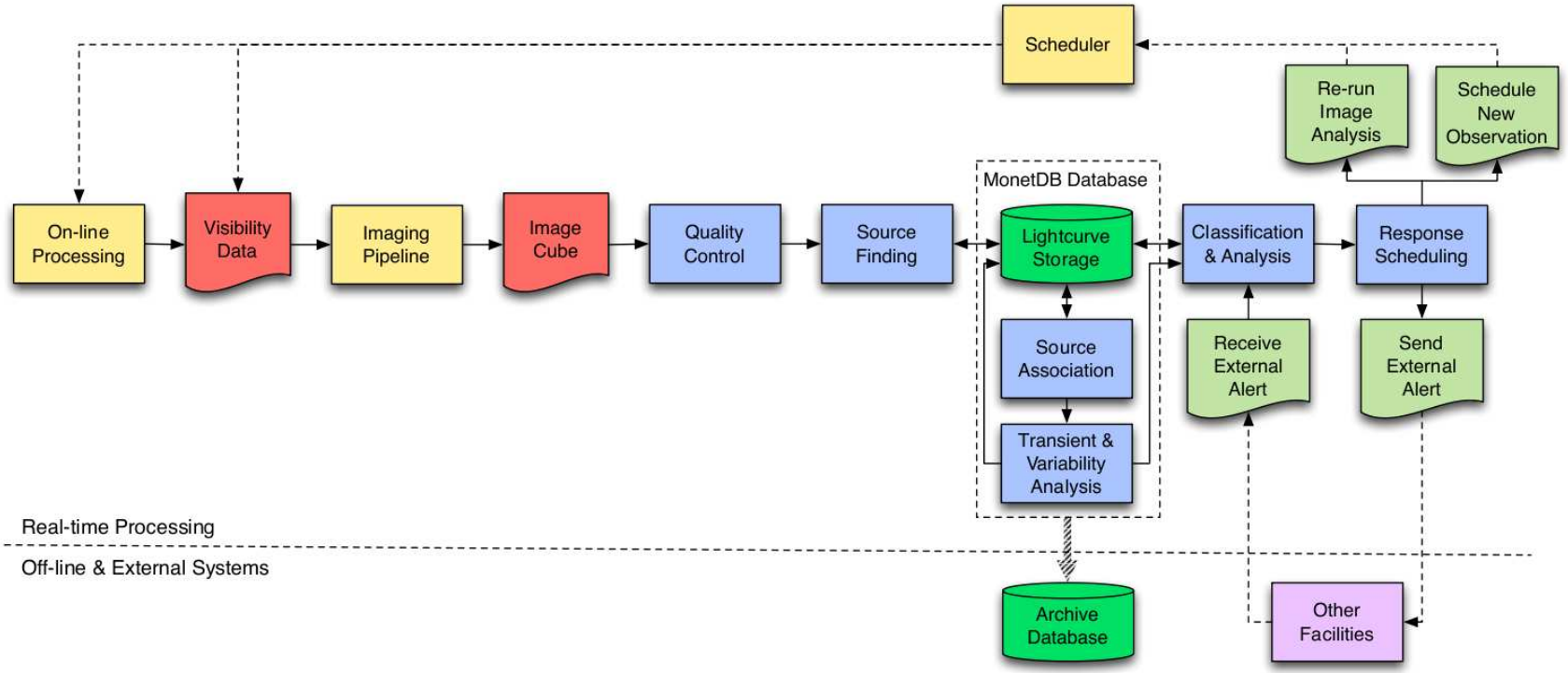
- Raw data $\sim 25\,\mathrm{TB/hr}$
- Distinct sources: $\sim 10^7 - 10^8$,
  - which are revisted many, many, many times
- Source properties reduce to $50 - 100\,\mathrm{TB/yr}$
- Peaks over 10,000 sources per second
- Automated Software Pipelines
  - Calibration/Imaging Pipeline
  - Transients Pipeline
- Actively use database $\Rightarrow$ move algorithms and statistics inside database engine
- Real-time data access, quick responses $\Rightarrow$ single node
- Accumulate data over time $\Rightarrow$ multiple nodes

LOFAR Status – The TKP Transients Pipeline



Courtesy: J. Swinbank

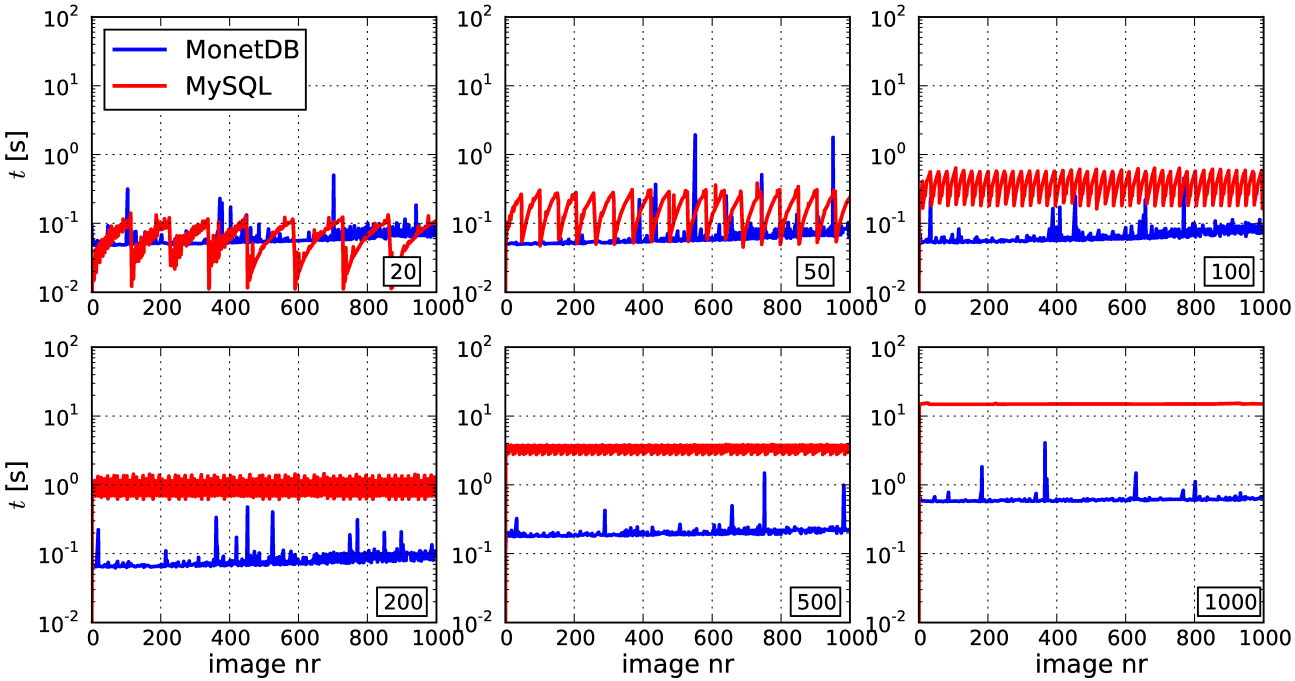
LOFAR Status – The TKP Transients Pipeline



Courtesy: J. Swinbank

- Move key-science frameworks more upstream in pipelines
  - Towards images or visibilities...
  - Array based query processing

LOFAR Status – Column-store Database

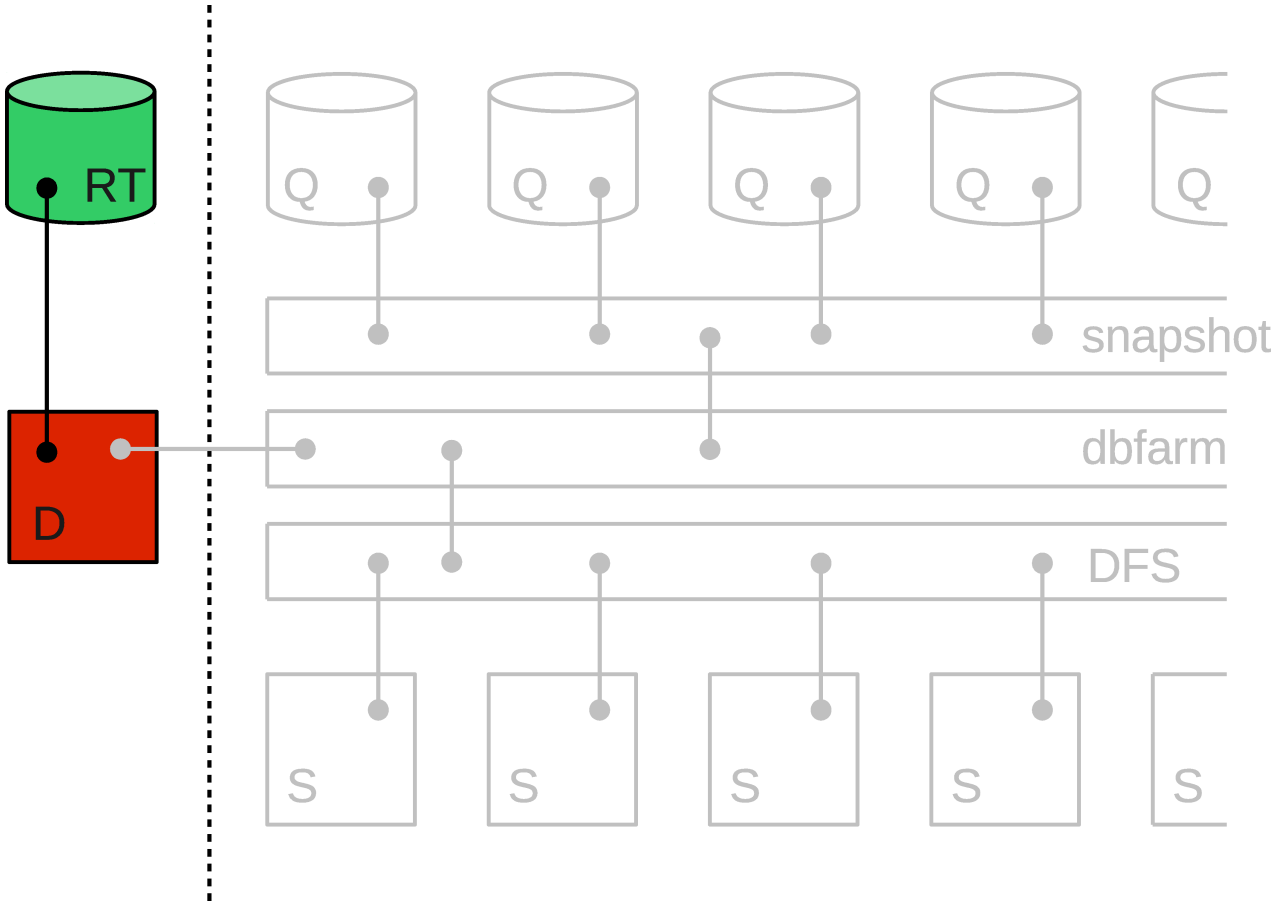


Processed a series of 1000 images (x axes), each containing the number of sources as labeled in the bottom right of the subplots.
Acc. response times of two most intensive queries shown on the y axes.

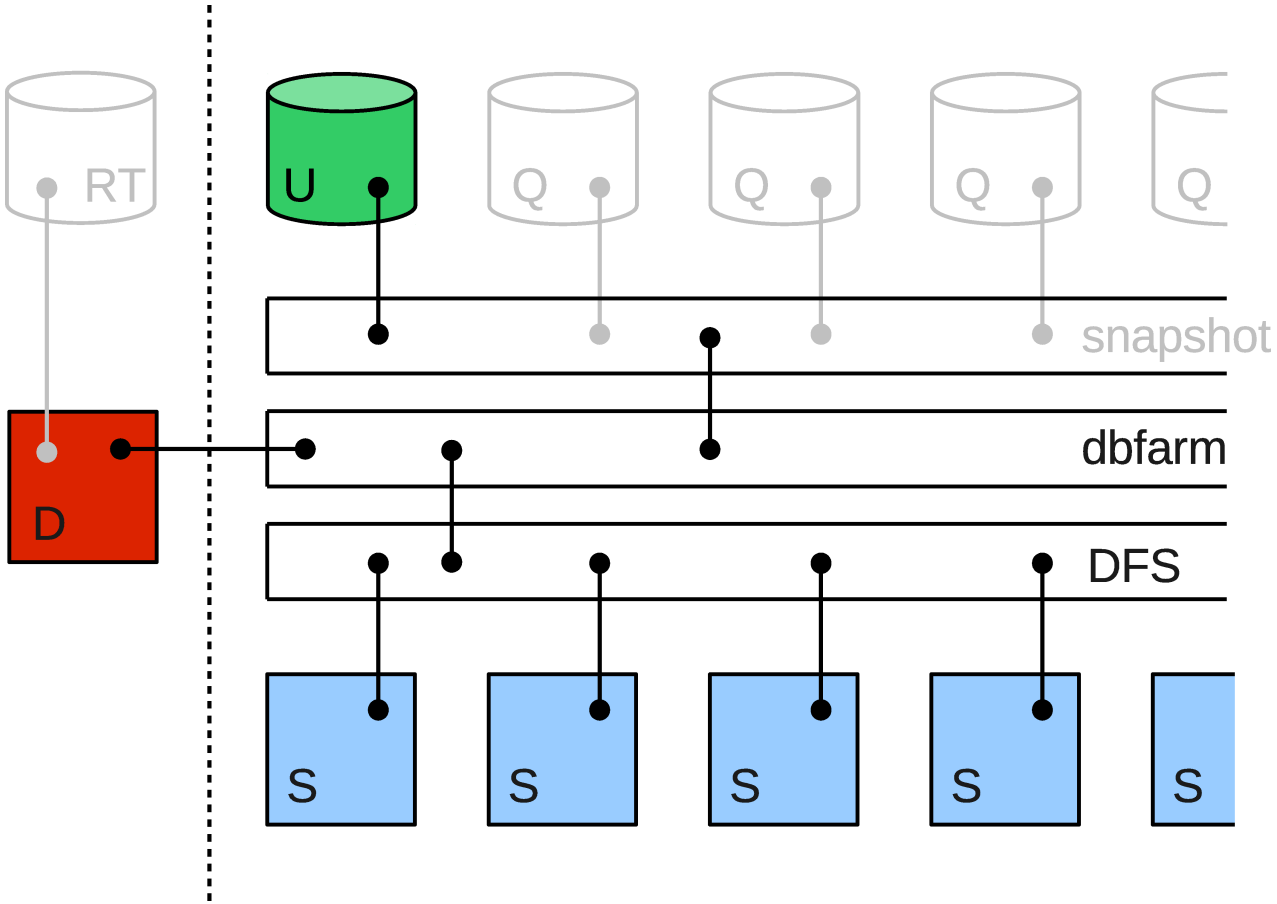
LOFAR Future Development Areas

- Consolidation of Transients Pipeline
- Continuing commissioning, processing first surveys
- Need for sky-tiling schema inside database
  - HEALPix, HTM, GIS
- Classification, ML, transient source model input
- Visualisation
- Scaling up

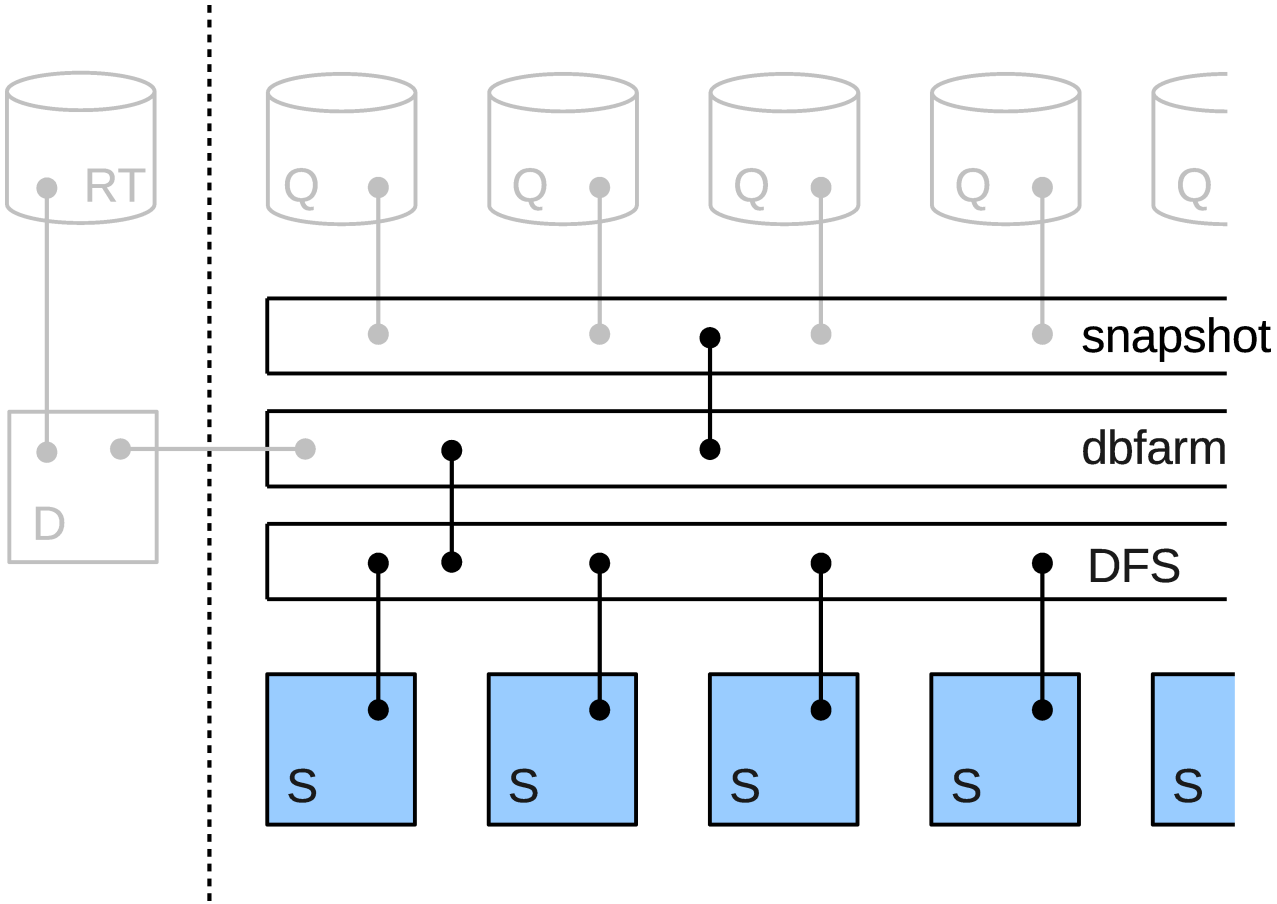
Distributed Light-curve Database, partition in time



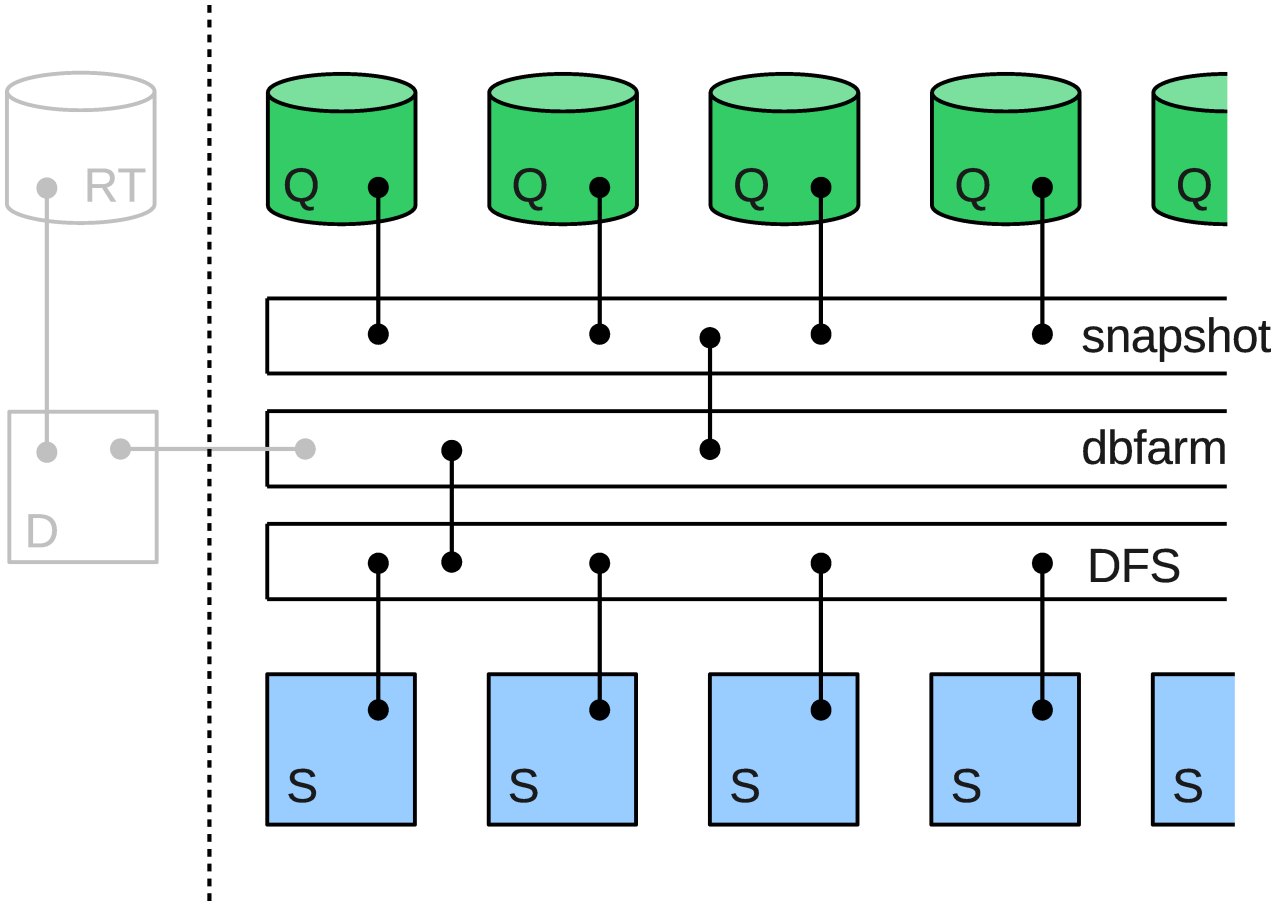
Distributed Light-curve Database, partition in time



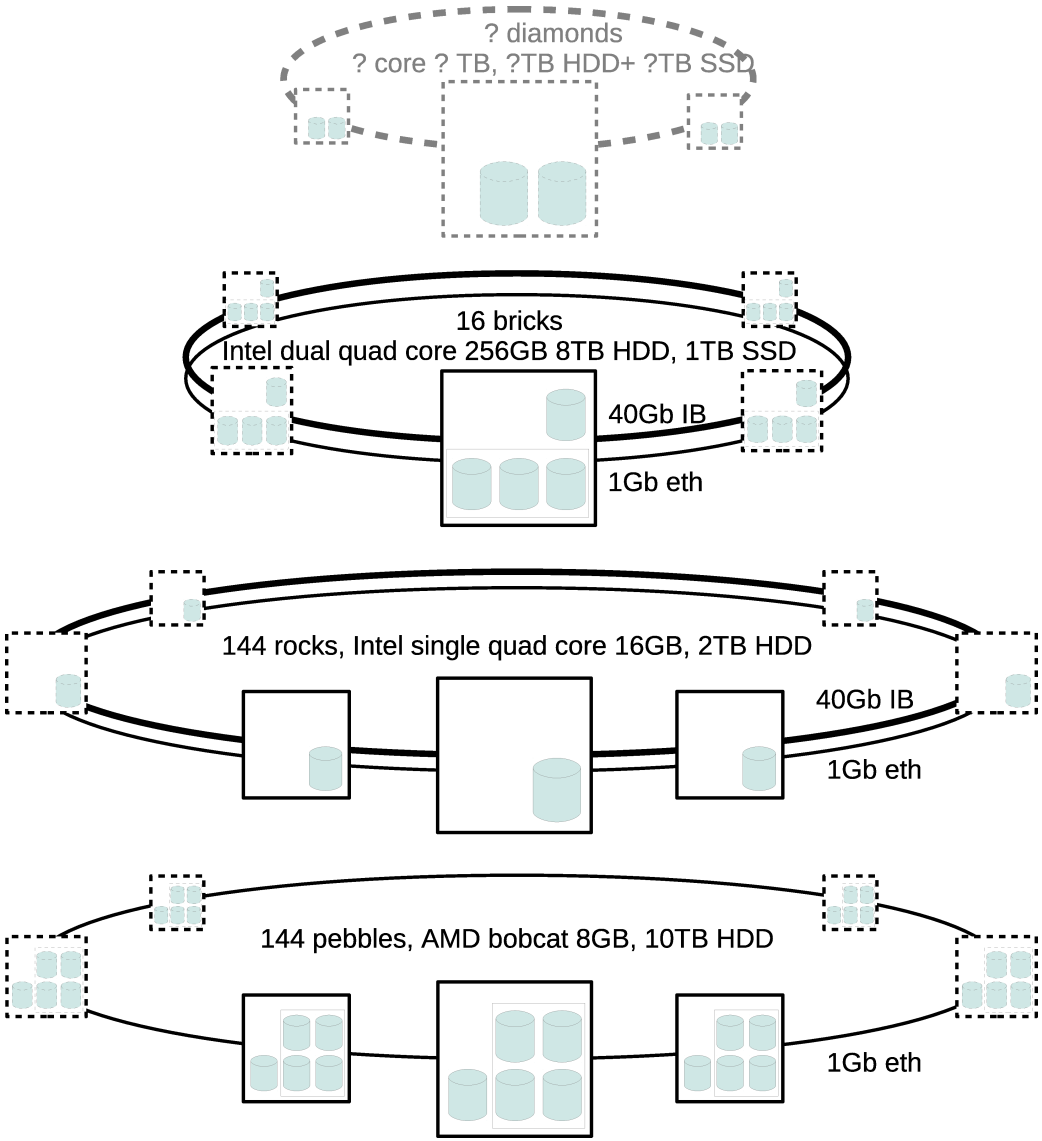
Distributed Light-curve Database, partition in time



Distributed Light-curve Database, partition in time



SciLens Platform, 300+ node experimentation cluster



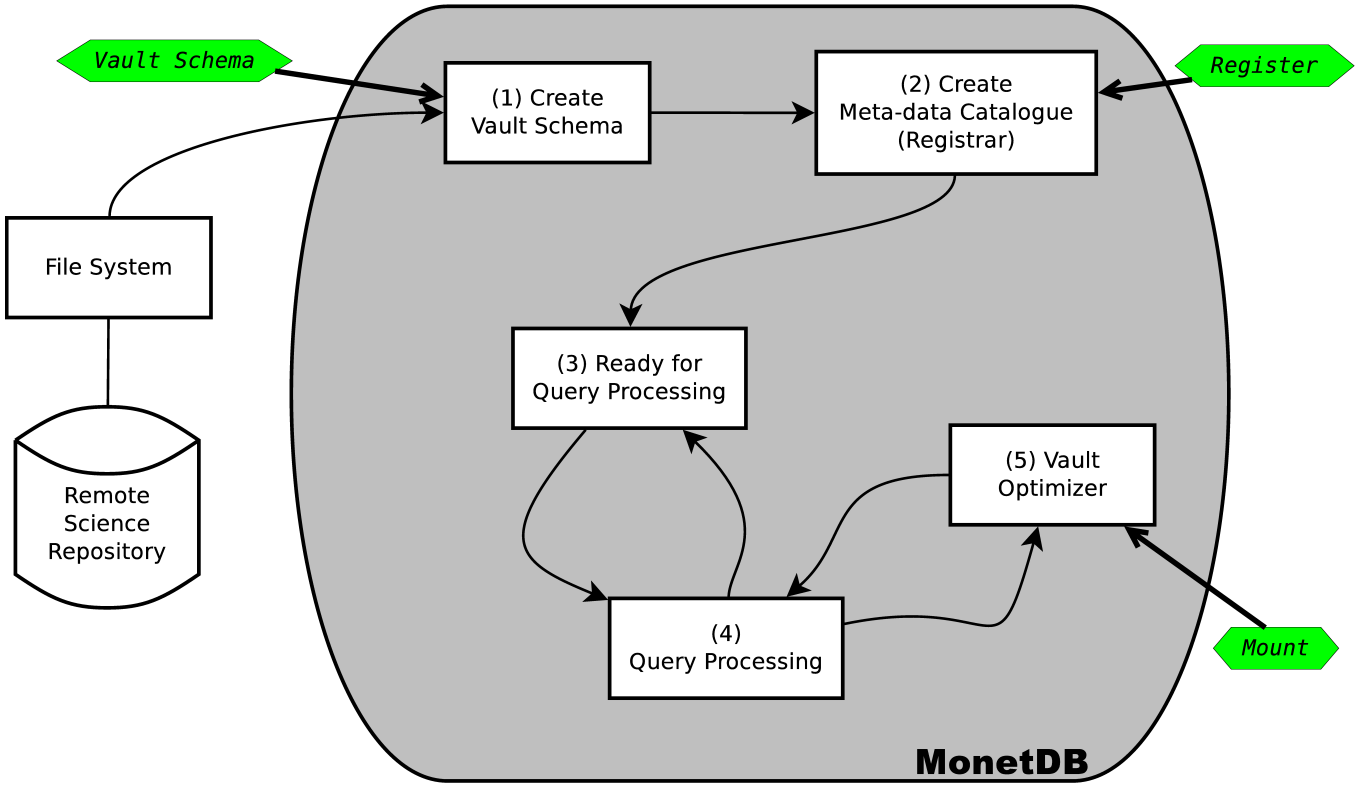
Enhancing data mining of PB dynamic catalogues
Array based (query) processing

- SciQL, backward compatible extension of SQL'03
- Symbiosis of relational and array paradigm
- Pushes operations to the data
- Fits well to multi-dimensional datasets, images & -cubes, visibilities
  - Enables detection framework to move upstream
- Advances data mining
  - Periodicity searches, FTs, cross- & autocorrelations
  - Multi-dimension transient searches
- More on `sciql.org` and youtube

Reprocessing raw/original data

- Storing is not a technical problem
- But retrieving, reloading, and therefore reprocessing is
- Data Vault framework couples dataset to database
  - initial load only metadata (compile time)
  - actual load at query time (optimized at execution time)
- Extends beyond csv files, (SEED), FITS, MS, HDF5
- Opening up repositories ?

Data Vault Framework



Courtesy: Y. Kargin

Simultaneous Multi-Messenger Observations

- Want to get light curves from external catalogues, when needed
- Synchronizing catalogues at different sites
- Data transport from remote locations
- Go for the low-hanging fruit, highest chance on success

Summary & Open Issues

- Column-stores boost performance
- Move key-science frameworks and statistics more upstream
- Distributed Databases
- Array-based Query processing
  - Aiding upstream processing
  - Advancing data mining
- Matching data formats to database for (re)processing
  - Data vaults framework
  - Open up repositories
- Synchronising catalogues at remote sites
- Create awareness for dynamic and distributed databasing