Online data analysis at the hard Xray micro/nano beamline P06.

J. Garrevoet
Deutsches Elektronen Synchrotron DESY, Hamburg, Germany

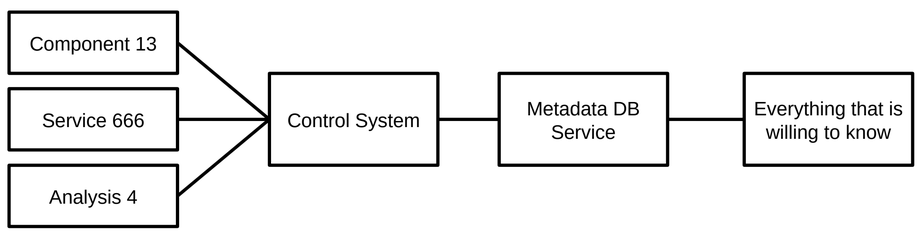
Motivation and Expectations and some complaining

In this day and age of more complex experiments and instruments, ever changing requirements and configurations, together with beamline users that are more disconnected from the technical side and methodological side offered by the beamline. It becomes a necessity to process the data for the users so they can solely focus on the interpretation of the results. Although this is not a requirement of any beamline nor is there any budget for efforts like this, it is only in the interest of the beamline to provide a service like this when one expects successful experiments and measurements which result in scientific publications.
In an effort to solve this issue, beamline manpower was alloted to this project, by reallocating tasks to other members of the beamline.

The expectations of the project were not to get real time data analysis, but online data analysis going. Providing users with chuncks of analysed data every couple of seconds.

Metadata, Metadata, Metadata, ….
Repeat after me!

Regarding Metadata is not only important to log all the experimental and sample information, but should also log the data acquisition settings, device settings, analysis settings, and so on. Not every value should be indexed so that it is fast to query it, but everything (and more) that one might need needs to be present.
The control system adds all the metadata to the scan by gathering it from all the various components.



There is an offline DB service available that uses a token for authentication and authorisation is done the same way as the access rights to the data on core file system.

Network Topology Manager

The entiry system is based on micro services that are distributed over multiple compute nodes. Service discovery is solved by the network topology service that is running in a fixed location. Upon starting a service, it will register itself with the topology manager so that other services can discover it on the network. The reverse happens when services are stopped.

Sink

The 3rd and final core component is the data sink. This is a service working in volatile memory space, keeping all the latest processed data streams readily available for any service that might be interested in the data.

The number of scans of which it can keep the data in memory is depending on the hardware and the used modalities. When memory becomes limiting, the oldest untouched dataset gets removed from the buffer to clear up memory space for fresh data.

Pipelines

Data pipelines can take many shape or forms. The easiest pipeline consists of 3 components: the pipeline master, the worker, and the sorter.

Pipeline Master

The master device has 2 main functions:
- Queueing events or data;
- Fan out to multiple workers.

The queueing aspect allows for the data to come in bursts that are higher than what your analysis chain might take.
To allow for a scalable approach, since demands change over time, it allows branching out to multiple downstream components.

Worker

The worker does whatever number crunching that needs to be done. The code that is required to include a new type of worker is just the number crunching part. Where the data is actually coming from is automagically determined by the framework.
The neat thing is that one can scale the number of workers dynamically to the current “load” coming from the beamline.

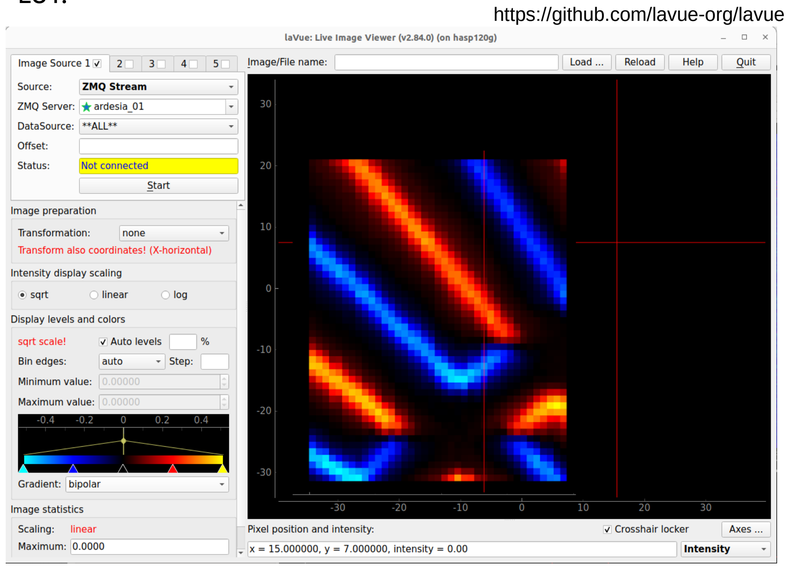
Sorter

The sorter is a fan in device to take in all the data that is produced by the workers and sort it again following their acquisition index since the received data can/will be out of order.
Besides the sorting it acts as a buffer to lower the amount of messaging events of the downstream components. This helps with the scalability of the entire system.

Data Visualisation

The visualisation of the processed data is published to the Lavue data viewer, which is developed at DESY. The available datasources are updated on the fly in case the data processing settings were adapted from one measurement to the next.
Automatic scaling of the visible range can be applied or the user can take full control of the histogram of the LUT.

https://github.com/lavue-org/lavue



Data Writers

Several data writer services are available that output the data in specific file formats or format the data for convenient viewing, such as graphs, images, …

Currently Supported Modalities

- Counters
- Positioning correction
  - Encoders
  - Interferometers
- XRF
- XRD
- Ptychography
- DPC
- XEOL
- XBIC
- XBIV
- Charge-carrier lifetime measurements.
- Tomography

Acknowledgements

The entire P06 team taking over tasks and for their input and patience!

Deployment

With automation and flexibility comes complexity. To keep the initial step towards operating the entire service low, most services are wrapped inside tango servers, which most beamline scientists know how to handle.
At P06, most services run on commodity hardware at the beamline, and consists of a little over 120 services.
Services that require more compute power or need to have access to dedicated hardware run on maxwell nodes in dedicated environments. The deployment on the Maxwell cluster has not been automated yet but currently does not form an issue due to starting and stopping of the service does not need to be done frequently. Priorities ;-)

Summary

The entire system has now been up and running for a while now and after some teething problems it has been running stable.
Users are happy with the online feedback they are getting and is often eliminating further processing on their side, making the time to publication shorter.