DATA ANALYSIS WORKBENCH

A. Götz , M.Gerring, O.Svensson, ESRF, Grenoble, France
S. Brockhauser, EMBL, Grenoble, France.

Abstract

Data Analysis Workbench is a new software tool being developed at the ESRF. Its goal is to provide a tool for both online data analysis which can be used on the beamlines and for offline data analysis which users can use during experiments or take home. The tool includes support for data visualization and workflows. Workflows allow algorithms which exploit parallel architectures to be designed from existing high level modules for data analysis in combination with data collection. The workbench uses Passerelle as the workflow engine and EDNA plugins for data analysis. Actors talking to Tango are used for sending commands to a limited set of hardware to start existing data collection algorithms. A Tango server allows workflows to be executed from existing applications. There are scripting interfaces to Python, Javascript and SPEC. The current state at the ESRF is the workbench is in test on a selected number of beamlines.

INTRODUCTION

The Data Analysis Workbench (DAWB) project has been started to create a platform for integrating data analysis programs. Data analysis programs usually exist as standalone programs. Each program has its own user interface, graphical and/or textual. Each program has its own internal architecture and choice of language and libraries. This makes it extremely difficult if not impossible to share code between programs.

Users are confronted by a myriad of different programs and interfaces to manage. Users generally revert to scripting languages to get around this problem. Scripting languages like Python allow them to provide a level of automation for running analysis programs. However scripts are usually not destined to be shared, documented or distributed. Scripts are also not well adapted to visualisation.

DAWB (see Figure 1) adds a layer on top of scripts to make data analysis and visualisation accessible to everyone and repeatable by introducing the concepts of a workbench, visualisation, scripting, workflows and experiment and data analysis design.

CONCEPTS

The goal of the Data Analysis Workbench is to provide powerful data analysis tools, both for online (synchrotron radiation beamlines) and offline data analysis. The combination in the workbench of a workflow tool (Passerelle, based on Ptolemy II) and a data analysis framework (EDNA) makes it straight-forward to develop advanced but robust data analysis workflows. Prototyping of workflows is facilitated by the workbench data visualisation support which makes it easy to visualise 1D and 2D data. The TANGO framework is used for online data analysis for both integrating beamlines hardware and for remotely controlling the workflow engine from the beamline GUI.

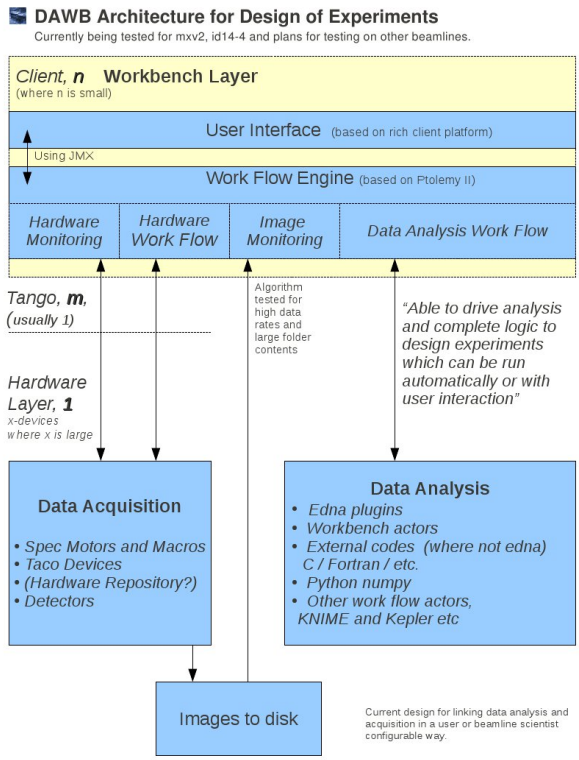

Figure 1: DAWB Architecture.

To achieve these goals the data analysis workbench is based on the following concepts:

Workbench

Data analysis is a sufficiently complex task that a simple approach of a main window with a few buttons is not sufficient. The Eclipse Workbench has been adopted as base platform for the data analysis workbench. The workbench can be best described by quoting the Eclipse documentation*: "*The Workbench aims to achieve seamless tool integration and controlled openness by providing a common paradigm for the creation, management, and navigation of workspace resources. Each Workbench window contains one or more*

* http://help.eclipse.org/indigo/index.jsp

perspectives. Perspectives contain views and editors and control what appears in certain menus and tool bars. More than one Workbench window can exist on the desktop at any given time."

Visualisation

Data analysis is a means to an end. In the initial, intermediate and final steps the raw and analysed data need to be visualised. One of the workbench's main goals is to offer a complete set of visualisation tools in 1D, 2D, 3D and the possibility of following changes as a function of time. Currently 1D and 2D are implemented. The aim is to offer viewers for all types of files typically used in data analysis in a single tool. Users need to learn one set of viewers for all data analysis applications called from the workbench.

Scripting

Scripting is an essential part of data analysis. It allows existing codes to be glued together rapidly and in a flexible manner. The workbench provides support for Python and Javascript as scripting languages. There is basic support for SPEC. Others languages might be added in the future. Python support includes a complete language development environment – PyDev[†]. Scripts can be called as standalone from a terminal or as part of a workflow.

Workflows

Workflows are a convenient way of programming and documenting the flow of data through algorithms. They can also be used to feedback to data acquisition systems when designing experiments. Workflows are naturally parallel and ease the programming of multiple execution threads. DAWB provides strong support for workflows, both online and offline. They are seen as a complement of scripts. Workflows are higher level constructs which assist the beamline scientist and user in building robust pipelines which use scripts.

Features

Import files as links - import large data directories into a project structure as links. This reduces file system overheads.

Open files in various formats - srs, dat, png, jpg, jpeg, tif, tiff, cbf, img, ciff, mccd, edf, pgm, cor, bruker, h5, nxs, pdb, (gz, bz2, zip)

Visualization - able to visualize data in 1D and 2D.

Workflow editing - viewing and editing of workflow data analysis pipelines based on Ptolemy 2.

Running of workflow - able to run workflows in a separate process which is controllable from the GUI and able to run without the user interface.

Eclipse projects - able to keep data in eclipse projects and analyze data further using user interface tools and python.

Drag and Drop - support of drag and drop of data from eclipse projects to Passerelle workflows.

Tango Devices - the monitoring of tango devices and the interaction with them via actors in the workflow.

Image Monitoring - the ability to monitor a directory of images using a thumbnail viewer, even if the directory contents are large and changing quickly.

HDF5 / Nexus - read and write hdf5 including slicing. Plotting of 1D and image datasets and slicing higher dimensions. Supported on a wide range of platforms and in multi-threaded environments. Nexus files in hdf5 also supported.

Data analysis - the support for python actors and EDNA actors natively and connections via system process actors for Fortran, C/C++ and other executable programs. EDNA is the most heavily used framework for integrating arbitrary data analysis to the workbench at the moment.

ECLIPSE/RCP

Eclipse/RCP[‡] is an essential part of DAWB. Eclipse/RCP offers a robust platform for building rich clients. It has many features which are useful for managing a workbench. The fact that it is a commercial quality open source product makes it easy to collaborate with commercial and non-commercial partners. The high quality documentation and large number of resources on the web make it easy to learn. Eclipse/RCP has enabled DAWB to be developed in a way that other institutes who have adopted Eclipse/RCP can use it even if they were not involved in the definition of or aware of DAWB.

WORKFLOWS

Workflows are relatively new in the synchrotron data analysis field. They are however widely used in other scientific fields like biology. They offer a higher level programming language than traditional textual languages e.g. Python, C and Fortran. Their goal is not to replace these languages but to complement them by providing a higher level tool for calling programs in these languages. For this reason workflows should remain high level and not become too fine grained. Too many nodes or graphs in a workflow quickly make it difficult to follow or maintain. Workflows like Passarelle in DAWB should not be confused with programming languages like Labview's graphical programming language G. In the case of the latter the goal is to do as much programming as possible in this language. It is not easy to interface it to scripting

[†] http://pydev.org/

[‡] http://www.eclipse.org/home/categories/rcp.php

languages like C or command line driven programs. DAWB on the other hand encourages users to delegate tasks as much as possible to high level modules (which could be scripts or standalone programs). The workflow is then dedicated to documenting the data flow and adding parallelism to the process. Workflows with the help of appropriate actors make accessing remote sources like web services, batch schedulers and cloud services transparent.

Workflow TANGO Server

The workbench is very good for developing workflows and running them when visualisation is part of the workflow. For some applications however once the workflow has been developed the workbench is not needed anymore. In those cases it is preferable to run the workflow "headless" i.e. without the graphical user interface. A TANGO device server has been written for this purpose. It implements start and abort commands plus the possibility of passing parameters to and from the workflow during its execution. It implements a state machine which reflects the status of the workflow's execution. The device server facilitates the integration of workflows in existing applications like MxCube [4].

MX workflow examples

Experiments performed at macro-molecular crystallography (MX) beamlines can to a large extent be automated [4]. However, the automation of data acquisition (beam delivery, sample changer, centring, mxCuBE) and data analysis (EDNA characterisation [5] and automatic data processing) have up until now been largely decoupled with only few possibilities of data analysis to influence data acquisition. New highly automated beamlines like MASSIF (ESRF upgrade program) will take advantage of advances in beam delivery and automation of sample handling. Together with recent advances in data analysis software packages this creates higher demands in term of overall automation of both data acquisition and data analysis.

The ESRF MX beamlines have successfully used DAWB for developing four workflows which significantly increases the coupling between data analysis and data processing:

Enhanced EDNA characterisation workflow for characterisation of MX crystals. This workflow can automatically re-adjust exposure time, oscillation width and detector distance in case these were not set optimally before collecting the reference images.

Crystal radiation sensitivity measurement workflow for precisely measuring the radiation damage susceptibility of crystals by sacrificing a crystal (or a part of a crystal).

Kappa goniometer re-orientation workflow (see Figure 2) for re-orienting a crystal to the desired orientation, e.g. aligning the principal unit axis along the spindle axis or aligning an even-fold symmetry axis along the spindle axis.

Mesh scan workflow for finding the part of a crystal which gives the best data by scanning in two dimensions a low-intensity beam over the crystal and at the same time evaluate the diffraction intensity.

A paper which describes these workflows in more detail has been submitted [3]. The workflows are planned to go into production later this year.

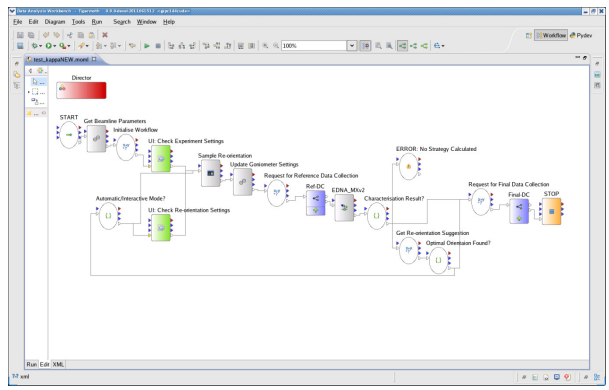

Figure 2: Workbench showing Kappa reorientation workflow for automatic characterisation of MX crystals

EXAFS workflow example

JESF is a Fortran code written for spectroscopy experiments which has been used for obtaining fast, first pass, data analysis at the ESRF. This code can be run online or with "zero dead time" attached to the detector using DAWB. At the ESRF, this works with shared memory. Since JESF uses a file and the speed requirements of the pipeline (2Hz maximum) are not currently large, we decided to write the data to a file and then run JESF on the file. However DAWB allows in memory operation as well. After JESF is run, the result files are opened in real time, so the plots refresh during the run. This pipeline has been tested with a SPEC macro writing to the shared memory currently as the beamline is not yet commissioned. DAWB allows this pipeline to be constructed / modified quickly and delivered to the beamline in a robust way. See Figure 3 for a screenshot of the JESF workflow.

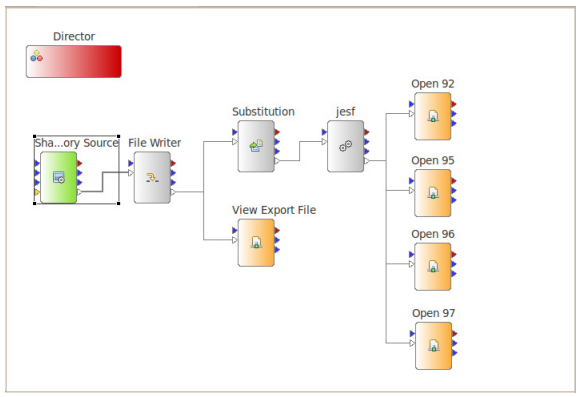

Figure 3: JESF Workflow for EXAFS online data analysis

Future Workflows

The capability of flexibly combining online-data-analysis actors with beamline control actors as well as reusing complete workflows as composite actors in a more complex workflow enables rapid implementations of new experiment protocols. These can include full 3D sample quality characterisation via diffraction tomography, evaluating multiple crystals in a raw sample to find the best diffracting crystal or even the best place and orientation of the best quality crystal. Using the already mentioned couplings of EDNA actors with the Data Collection composite actor, like Radiation Sensitivity Workflow, or Enhanced Characterisation Workflow, even more sophisticated workflows can be built e.g. chaining them one can gain empirically a susceptibility coefficient in the first step and then immediately use it during the characterisation for more precise results. Chaining basic calibration workflows one can automate complete beamline checking protocols. The list of possible workflow programming is endless and comparable to the list of scripts. To gain the most out of workflow design, it is important to keep the structure of the workflows clean and easy-to-understand (comments on the canvas if needed and documentation of course) as well as defining a proper data model to be used all along the communications between the actors, so that the dataflow is kept clean.

COLLABORATION

DAWB is an Open Source program and welcomes collaborators. The source code is on a publicly accessible repository[§]. The source code is released under a modified GPL[**] licence which allows actors developed by third parties not to be under GPL. Workflows developed with DAWB are not bound by the DAWB licence. They can be released under the licence of choice of the developers. Open source licences are encouraged because they allow other developers to study the source code and improve it and or learn from it.

CONCLUSION

DAWB has demonstrated that it is possible to provide a commercial quality workbench for synchrotron data analysis which supports workflows, visualisation and scripting for online and offline data analysis. The first examples of workflows running on MX and EXAFS beamlines are very promising. We think that in the future workflows are going to be an essential part of managing high-level scientific data analysis. DAWB has the features necessary to achieve this. We encourage others who are interested in scientific workflows either as developers or as users to give DAWB a try[††] and if interested to join the DAWB collaboration. We think we should collaborate on the core tool but leave the choice to compete or collaborate on workflows open.

ACKNOWLEDGEMENTS

We would like to acknowledge the following groups: iSencia[‡‡] for providing the Passerelle workflow tool under an open source licence and for doing the Eclipse/RCP version, Soleil for introducing us to Passerelle, Diamond for making parts of GDA and SDA available under an open source licence.

REFERENCES

[1] Johan Eker et. al (2003). “Taming Heterogeneity - The Ptolemy Approach” in Proceedings of IEEE, vol. 91, no. 1

[2] Gwenaëlle Abeillé, Majid Ounsy, Alain Buteau, “A Graphical Sequencer for SOLEIL Beamline Acquisitions”, ICALEPCS 2007 Proceedings

[3] Sandor Brockhauser et. al. “The use of Workflows in the design and implementation of Complex Experiments in Macromolecular Crystallography””, to be published in Acta D

[4] Antonia Beteva et. al. (2006). “High-throughput sample handling and data collection at synchrotrons: embedding the ESRF into the high-throughput gene-to-structure pipeline”, Acta Cryst. D 62, 1162-1169.

[5] Marie-Francoise Incardona, et. al. (2009). “EDNA: a framework for plugin-based applications applied to X-ray experiment online data analysis”, J. Synch. Rad. 16, 872-87

[§] http://code.google.com/a/eclipselabs.org/p/dawb/

[**] http://www.gnu.org/copyleft/gpl.html

[††] http://dawb.org

[‡‡] http://www.isencia.com/