SAM, IFDATA, Metadata and ART Integration

Adam Lyon, Robert Illingworth, Marc Mengel, Andrew Norman
2 May 2012

# Introduction

ART is a common analysis and reconstruction framework written by the CS/CET group and used by NOvA, mu2e, g-2, and LBNE. The main purpose of ART is to read input event data files, process them with some algorithms, and write the results to output files. The problem that this project addresses is integrating the file delivery system with ART for delivery of the input event data files as well as injecting metadata into the output files necessary to declare such files into the SAM catalog.

The system for file deliveries will be SAM, used very successfully at the Run II experiments. For Run II, the SAM team had to create an integration layer between SAM’s C++ API and the experiment’s framework system. Writing this integration layer was difficult, time consuming, and introduced unwanted dependencies. To mitigate these problems, REX/DH introduced SAMWeb.

SAMWeb consists of an http service layer that sits in front of the SAM client. SAMWeb accepts https calls, translates them into SAM commands (for the python API), and returns the results, again over http.[[1]](#footnote-1) With SAMWeb, the framework communicates with SAM over http using a standard http communication library and needs no SAM specific code. The SAM team needs no intimate knowledge of the framework and the ART team needs no intimate knowledge of SAM. A document describing the interactions between SAM and ART is at CD-docdb-4684 and is not repeated here.

In the course of meeting with the CET group, we note that there is also a need to integrate the IFDATA system (see its Project Plan elsewhere) into ART. SAM is responsible for discovering and delivering the data files to some location accessible to the job’s worker node. Retrieving the files from that location to the worker node itself is handled by the IFDATA system. For example, IFDATA can move files on local disk, NFS disk, and via GridFTP or SRMCP to the proper spot on the worker node (doing this avoids a requirement of having SAMCache on each worker node). The integration of IFDATA into ART is covered by this project.

Files are cataloged in SAM using metadata. For Run II, separate metadata files were produced by the applications and were used along with the data file for storage into SAM. These separate metadata files were a pain to keep around and in many cases were not produced correctly by the applications. ART has the ability to store a small database in the data file itself, and in fact ART uses it for persistent provenance information. If we could store SAM metadata in that database as well, we could have self-describing data files. Self describing data files would be a huge simplification in how we import such information into SAM and would lead to ease-of-use improvements with users. There was an initial requirements document written and available as CD-docdb-4627. The contents of that document will not be repeated here. We had a meeting with the CET group on March 13, 2012 to go over the details. The CET group is implementing this part of the project.

For now, NOvA is the main customer for this project.

# Tasks

Below are the tasks for the project. Many of them are external.

## Determine requirements for SAMWeb + metdata and ART integration

This task was previously performed and resulting in CD-docdb-4627 (requirements document for metadata) and CD-docdb-4684. The latter document is more about implementation, but is derived from a meeting where requirements were identified and discussed.

## Determine the SAMWeb and ART interactions

This task involves determining what SAMWeb calls ART needs to make and when to make them (e.g. when to call get next file and what to do with the response). This work was completed at a meeting on March 6, 2012 attended by REX/DH and the CET groups and was written up by Jim Kowalkowski and Adam Lyon in CD-docdb-4684.

## Determine the metadata and ART interactions

This task involved the behavior and life cycle of metadata for the output files. We had a meeting on March 13, 2012 and worked out the details. The CET group has notes embodied in that they are following for the implementation. See ART Feature request 2397 at https://cdcvs.fnal.gov/redmine/issues/2397

## Code up the SAMWeb calls into ART

This task involves constructing an Art module to add the file delivery functions with SAMWeb as described by CD-docdb-4684. It is being performed by the CET group.

I estimate the work should take one month to complete and it is already underway. While it can be coded concurrently with the next task 2.5, testing cannot occur until 2.5 is completed. There is also a dependency on the ART metadata handling system covered in a separate project.

## Add an IFDATA ART Service

As mentioned above, IFDATA needs to be integrated into ART in order for ART to transfer the file to the worker node. REX/DH will write an ART service to provide this functionality. This IFDATA Art Service will live in its own ups product to avoid introducing direct coupling between IFDATA and ART. Marc and Adam are working on this task. I estimate two weeks to complete.

## Code up the metadata system into ART

The CET group will code up the metadata system into ART in accordance with Feature request #2397. I estimate one month of work for this task to be accomplished.

## Write a SAM declare script

REX/DH (Marc Mengel) will write a script that will take a data file with an embedded SAM metadata DB, extract the metadata needed by SAM, and perform the SAM declare placing the file into the catalog. The CET group is providing us with an example metadata extractor. I estimate two weeks for this task.

## Test the system

Once all of the pieces are complete, the system must be tested. Testing should be lead by REX/DH (e.g. Marc Mengel) but will need contributions and analysis from the CET group, Andrew Norman, Adam Lyon, and Robert Illingworth. I estimate one month to do testing and fix problems.

## Deployment

Deploying this system is quite simple. CET needs to release ART with the additional functionality. REX/DH needs to release a relocate-able ups compatible tar file that the CET can incorporate into there “externals” release. Lynn Garren typically does this packaging for CET. I estimate a day for this work. Then it is up to NOvA to upgrade to these releases.

## Integrate with Jobsub scripts

Part of the IFDATA project, introduces a coupling here.

The total time for this project is about 3-4 months.

# Risks

Below are risks we need to watch.

## Reliance on CET

The main task, coding SAMWeb calls into ART, is performed by the CET group. This group has a heavy load and our project may fall behind in favor of higher priority tasks.

## Coupling to other projects

This project is coupled with the IFDATA project and cannot be considered complete until that project is complete as well.

## Rejection by customer

If NOvA decides to use another data handling system, this project is largely unnecessary. Though integration with IFDATA would still be useful.

# Conclusions

When this project is complete, NOvA will be able to run jobs that can retrieve data via SAM natively with ART. Using SAMWeb makes deployment virtually trivial and eliminates a direct coupling between the ART framework and SAM itself. Metadata will injected into the output files and those files will be self-describing.

It will be easy to then bring this system to other ART-using experiments.

Future work we should think about is how the output files will be returned back to the user.

1. SAMWeb’s functionality is described at https://cdcvs.fnal.gov/redmine/projects/sam-web/wiki [↑](#footnote-ref-1)