The SLAC Workflow Engine and Data Catalog

Brian Van Klaveren (SLAC National Lab)

The HEP Software Foundation Workshop
2015-01-21

Introduction

The SLAC Workflow Engine and Data Catalog are middleware solutions for data processing and data handling. Both systems are generalized, though a modest amount of work would be needed to make them distribution-ready.

- Originally developed to meet the needs of the Fermi Gamma-ray Space Telescope
  - Needed to handle Level 1 processing, MC, user jobs, etc..
- They are designed to be independent components
  - i.e. NOT inextricably coupled, nor part of an overarching framework
  - Integration achieved through plugins
  - Enhanced functionality also achieved through plugins

Workflow Engine (aka “Pipeline-II”)

Pipeline-II is a Workflow Engine designed to execute user-defined **D**irected **A**cylclic **G**raph workflows composed of batch jobs and user scripts to aid in orchestration.

The workflow support fan-out and fan-in.

It is not tied to any specific processing system.

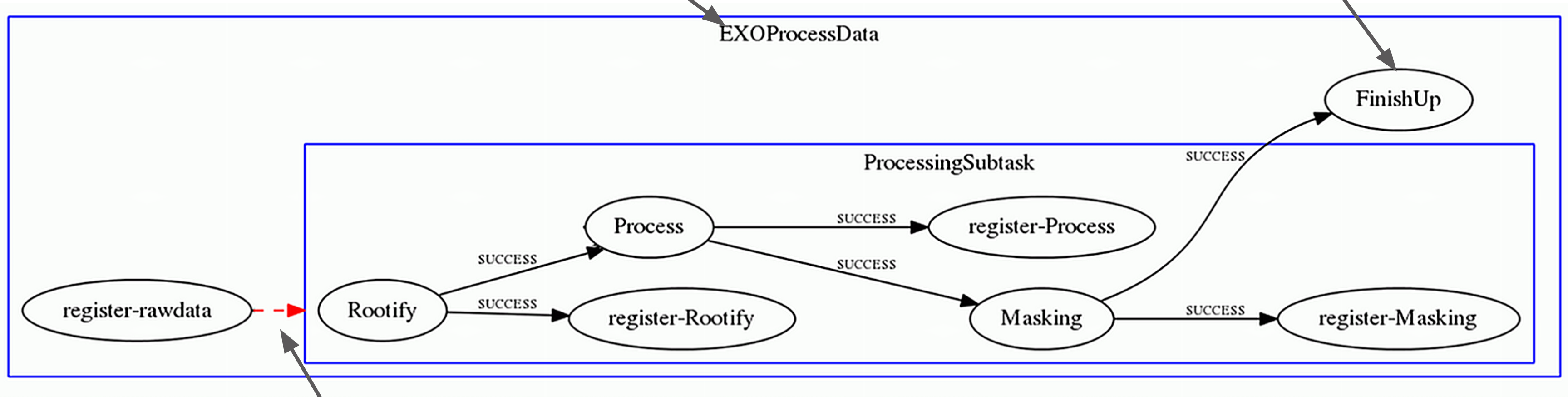
It’s basically a meta-scheduler...

- Can submit to a variety of processing systems
  - Jobs have been submitted to LSF, PanDA, DIRAC, NERSC, HTCondor, LSF, Torque/QSUB, SGE, Amazon/Cloud
  - If it can behave like Batch system, you can submit to it
- Centralized job reporting
- Web Interface

DAG Workflow with fan-out/in

Task

Process



Process sets up and spawns *n*-many streams of task ProcessingSubtask

The Data Catalog

The Data Catalog is a globally addressable metadata database for your files.

That metadata is stored in a virtual hierarchy, so it looks like a file system as well.

It is not tied to any specific processing system.
It is not tied to any file system or protocol.

As a Service

- Web Interface
- CLI interface
- Plugin for Workflow Engine
- Crawler validates registration and optionally extracts metadata asynchronously from registration
  - Both file metadata (i.e. size) and data metadata (i.e. run id, start time, event count, etc...)

In Use:

  - Fermi Gamma-ray Space Telescope (> 5 years)
  - EXO, LSST-Camera, as well as people in smaller projects
  - Working with LSST-Data Handling, Heavy Photon Search, and others

Actively being rewritten around RESTful API

- Implementation is DB agnostic thanks to VFS layer
  - Implementations for MySQL, Oracle, HSQLDB, PostgreSQL. No barrier to using NoSQL
- Supports ACLs on records.
  - Pluggable auth. Investigating LDAP integration
- Reworking web interface
- Need to finish/refine client libraries
  - “Remote” by default
  - Support Python first, then Java and Javascript

In summary

- Our Workflow Engine is developed around computing requirements.
  - ...but not explicitly physics
- Our Data Catalog is developed around data handling and organization
  - ...but also not explicitly physics
- Both applications are loosely-coupled and can run standalone
  - However, there is power in packaging
- Use of these tools has been adopted by diverse experiments
  - We tend to cater towards smaller and medium size experiments
  - Increased adoption drives generalization and increases quality

We would like more people to use our tools!