A Prototype Data Processing System for JDEM

Eric H. Neilsen, Jr. & Erik E. Gottschalk
Fermi National Accelerator Laboratory

The Joint Dark Energy Mission is a collaboration between NASA and the U. S. Department of Energy to build a space telescope to study dark energy using several complementary techniques.
Analysis of the data from this mission will require the uniform, reproducible, and reliable execution of a complex set of data reduction applications.
We present a prototype data processing infrastructure created in preparation for developing the infrastructure needed to host JDEM data processing operations.

Goals for the JDEM Prototype

Provide a context for testing elements of a bulk data processing infrastructure.

Create a baseline against which candidate production components and architectures may be measured.

Data Processing Infrastructure Components

Processing pipeline
Reference application: mtpipe

We used the SDSS mtpipe pipeline because it is stable, familiar to the developers, and implements processing steps typical of pipeline reduction of survey data. The SDSS data processing group did not automate mtpipe processing to the same degree as it did other SDSS pipelines, so, unlike these other pipelines, creation of an data processing system was not a "solved problem."

Quality control and monitoring
Reference application: Pipeline produced files and diagnostic output

The sample application, mtpipe, generates an assortement of plots and text files that can be used to find and diagnose problems. mtpipe includes several utilities for further exploration of the data.
While this approach is been successful in many projects, including the execution of mtpipe for SDSS, significant development effort might be saved using a common infrastructure for monitoring QC data.

Bookkeeping and provenance
Reference application: Standardized directory structures, file names, and FITS headers

Data files produced by mtpipe incorporate metadata describing their provenance, and the scripts executed in the processing jobs store the data in directories and files named according to this metadata.
Both the production SDSS execution of mtpipe and likely candidate systems for JDEM will store metadata in a database. Such a database will be included in the prototype in the future.

Distributed processing job management
Reference application: globus & condor

Globus and Condor allocate compute resources and manage jobs on Fermilab computing resources.

Mass storage
Reference application: SRM, dCache & enstore

Enstore provides a convenient interface to data on tape, dCache caches stored data on disk and SRM provides a versatile external interface to the data.
The production SDSS system, and any likely candidate system for JDEM, stored a subset of the data in a database; such a database will be included in future versions of the prototype.

Workflow management
Reference application: DAGman & custom scripts

Custom scripts use detailed knowledge of the data to generate Directed Acyclic Graphs that describe the required processing steps and their dependencies; DAGman executes these steps using Condor.
The data processing operations that will be required by JDEM are likely to be much more complex than those required by the prototype sample application; the JDEM infrastructure will benefit greatly from higher level workflow design and management tools.

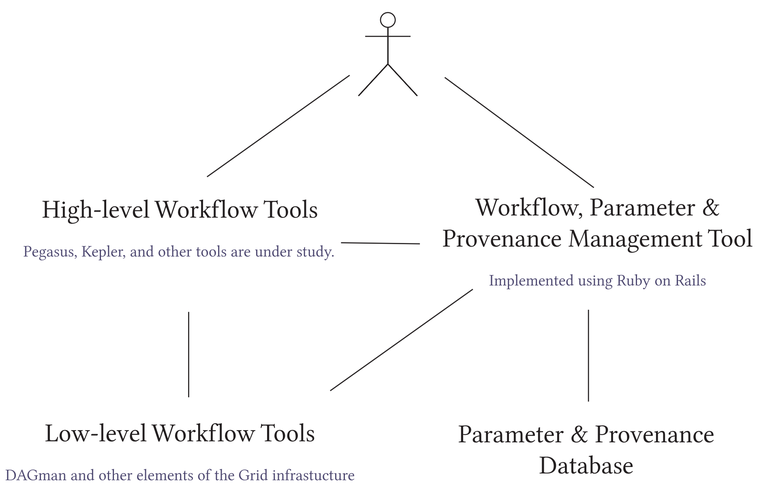
Summary

Significant institutional infrastructure is needed to run a proper data processing factory. Fermilab's existing infrastructure provided the components used in this prototype. Given the processing application itself and this infrastructure, very little additional development was required. Other experiments at Fermilab that use this infrastructure show that it can scale to JDEM size projects or larger. Features not yet supported by the prototype, such as a database to store data and metadata, must inevitably require some development to tailor it to the relevant data. The continuing evolution of Fermilab's infrastructure and adoption of new technologies will reduce the overall development and operations effort required while improving the quality of the processing pipeline as a whole.

Research & Development

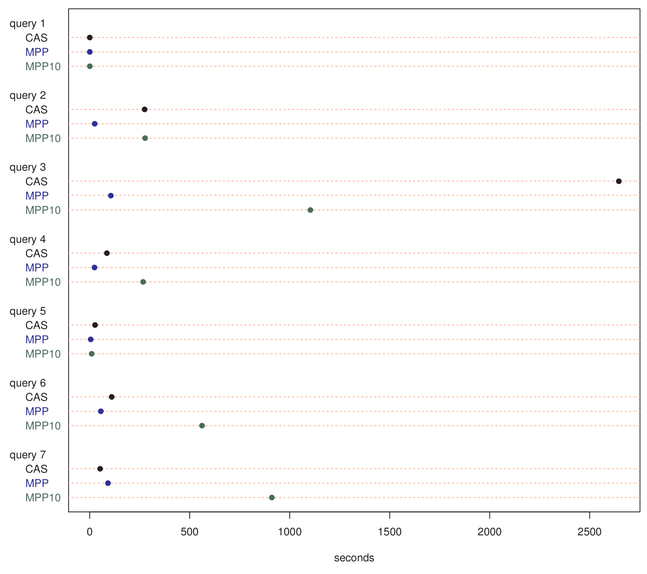
LatticeQCD Workflow, Bookkeeping and Provenance

Fermilab, Vanderbilt University and the Illinois Institute of Technology are developing a workflow, bookkeeping, and provenance tracking system in support of the SciDAC Lattice Quantum Chromodynamics project. Developers from the Lattice QCD project and JDEM are collaborating to make this infrastructure as general as necessary to support JDEM data processing as well.
This system will simplify the specification and automation of the data processing workflow.



Database Storage and Access to Data

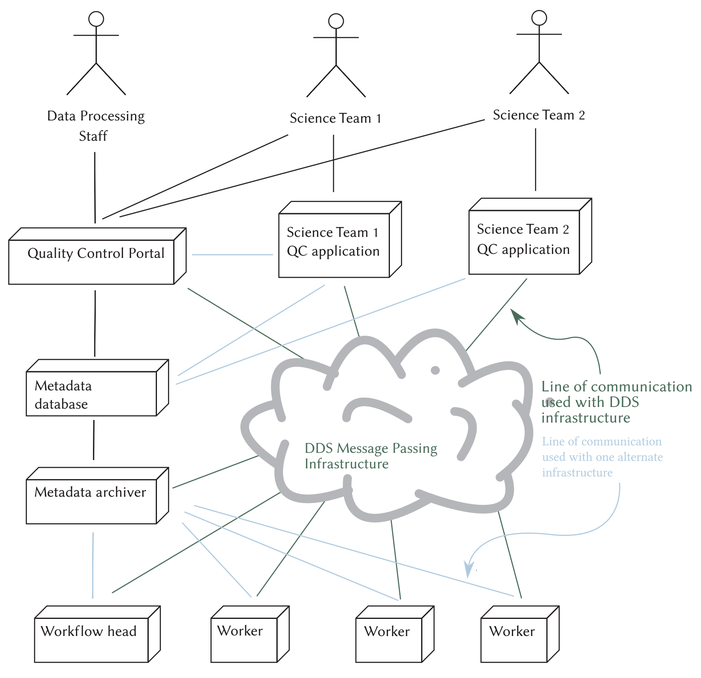
Both the data processing pipeline itself and collaboration users will need to execute complex queries on the data produced by the pipeline; some or all of the data will need to be stored in a database in addition to being archived as files. The expected data volume, between 10 and 100 terabytes, presents problems for many current databases; optimizing and configuring such a database will require significant effort.
The large size and expected usage patterns suggest that a massively parallel processing (MPP) database may be significantly easier to develop. We are therefore testing one such database, Greenplum. (Greenplum is a commercial product based on MapReduce and PostgreSQL.)



We ran a set of test queries on three times: once on the SDSS Catalog Archive Server (CAS), a carefully tuned "traditional" relational database, once (MPP) on an MPP database with a copy of the SDSS tables contained in the CAS, and once (MPP10) on the MPP database with these data replicated ten times to simulate a larger database.
The lone query on which the CAS performed better used a cone-search to cross match two catalogs. On this query, the CAS had the advantage of using hierarchical triangular mesh (HTM) position indexes, while the MPP query was did not use indexed positions. Implementation of HTM or another more advanced mechanism for cone searches is expected to be practical on the MPP database as well, if it is necessary, but even without it the MPP database took less than twice the time the equivilant query took on the CAS.

Publish-Subscribe Quality Control Infrastructure

Immediate access to data quality metadata will be of interest not only to the data processing staff, but also to the science teams and other collaborators. Different parties will be interested in different sets of data, and are likely to prefer to view it in different ways.
We are therefore considering a publish-subscribe architecure for distribution of quality control data. Such an infrastructure would allow different users to customize the data received and develop specialized quality control applications.
We are evaluating a data distribution infrastructure based on the Data Distribution Service (DDS) standard.



References and Further Information

JDEM: http://jdem.gsfc.nasa.gov
DAGman: http://www.cs.wisc.edu/condor/dagman/
Condor: http://www.cs.wisc.edu/condor/
Globus: http://www.globus.org/
LQCD Workflow: http://cd-docdb.fnal.gov/cgi-bin/ShowDocument?docid=2988
SRM: http://computing.fnal.gov/ccf/projects/SRM/
dCache: http://www.dcache.org/
enstore: http://www.fnal.gov/docs/products/enstore/intro.html#46888
mtpipe: Astronomische Nachrichten, Vol.327, Issue 9, p.821
SDSS CAS: Computing In Science and Engineering, Vol.10, Number 1, p.30
PostgreSQL: http://www.postgresql.org/
MapReduce: http://labs.google.com/papers/mapreduce.html
Greenplum: http://www.greenplum.com/
DDS: http://portals.omg.org/dds

Fermilab is operated by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the United States Department of Energy.