Astronomical Data Analysis Software and Systems XIX
ASP Conference Series, Vol. XXX, 2009
Y. Mizumoto, K.-I. Morita, and M. Ohishi, eds.

FERMILAB-CONF-09-631-CD

A Prototype Data Processing System for JDEM

Eric H. Neilsen, Jr.

Erik E. Gottschalk

Fermi National Accelerator Laboratory

Abstract. Fermilab is developing a prototype data processing system for the Joint Dark Energy Mission (JDEM) Science Operations Center (SOC). The SOC requires tools for systematic and timely execution of computing jobs; data management; and systematic record keeping for provenance, workflow organization, and quality control. Drawing on our experience in processing large volumes of data, we are constructing a prototype system using existing infrastructure and a variety of tools, including distributed computing and mass data storage infrastructure at Fermilab. This system provides an environment within and against which potential architectures and components of the production system can be evaluated. We describe the prototype and its uses.

1. Introduction

The Joint Dark Energy Mission[1] is a collaboration between the U. S. Department of Energy and NASA 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.

2. Goals

The prototype system will serve two purposes: first, to provide a context in which to test candidate technologies for use in a production system, and second, to provide a baseline against which such technologies can be compared.

3. Baseline Components

The initial construction of the prototype system is based entirely on components that are currently used at Fermilab. These components serve as references against which alternatives can be compared, and collectively provide a context in which tests can be made.

[1] `http://jdem.gsfc.nasa.gov`

The first such element is the processing pipeline itself. We use the SDSS MTPIPE pipeline (Tucker et al. 2006) 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 a data processing system was not a "solved problem." This part of the prototype will be replaced by applications useful to JDEM as they are developed.

Another element is the mechanism for monitoring and quality control. The sample application, MTPIPE, generates an assortment 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 has 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 are essential in a large scale data processing infrastructure. 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.

A distributed processing job management system manages jobs on available computing resources. The baseline system uses the standard tools used at Fermilab: Globus[2] and Condor[3].

Infrastructure to provide access to mass storage is also required. For the baseline system, Enstore[4] provides a convenient interface to data on tape, dCache[5] caches stored data on disk and SRM[6] provides a versatile external interface to the data. The production SDSS system, and any likely candidate system for JDEM, store a subset of the data in a database; such a database will be included in future versions of the prototype.

Finally, a workflow management system is needed to organize the processing steps and initiate them as appropriate. In the prototype system, custom scripts use detailed knowledge of the data to generate directed acyclic graphs that describe the required processing steps and their dependencies; DAGman[7] 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.

---

[2] `http://www.globus.org`

[3] `http://www.cs.wisc.edu/condor`

[4] `http://www.fnal.gov/docs/products/enstore/intro.html`

[5] `http://www.dcache.org`

[6] `http://computing.fnal.gov/ccf/projects/SRM`

[7] `http://www.cs.wisc.edu/condor/dagman`

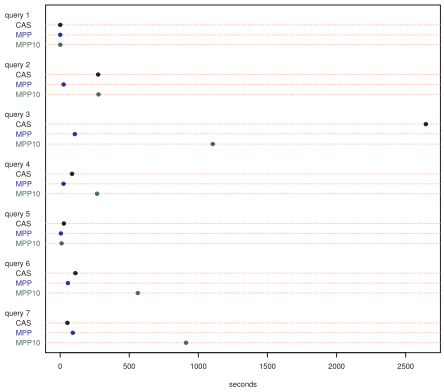

Figure 1. Comparison between SDSS CAS, an MPP database, and an MPP database with 10 times as much data.

4. Current Development and Evaluations

4.1. 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.

4.2. Data 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 have tested one such database, Greenplum. (Greenplum[8] is a commercial product based on MapReduce[9] and PostgreSQL[10].) For results of the test, see Figure 1.

We complete three executions of a set of test queries: once on the SDSS Catalog Archive Server (labeled "CAS" in Figure 1), a carefully tuned "traditional" relational database; once on an MPP database (MPP) with a copy of the SDSS tables contained in the CAS; and once on the MPP database with these data replicated ten times (MPP10) 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

[8] `http://www.greenplum.com`

[9] `http://labs.google.com/papers/mapreduce.html`

[10] `http://www.postgresql.org`

hierarchical triangular mesh (HTM) (Barrett 1995, Kunszt et al. 2000) position indexes, while the MPP query 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 equivalent query took on the CAS.

4.3. 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 architecture 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[11].

5. Conclusion

The computing tools and infrastructure available at Fermilab provide many important elements necessary for processing data from a large, data intensive experiment. A simple system constructed from these elements is a valuable tool for evaluating candidate technologies for use in a true production system and for identifying areas where additional development is required.

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

References

Barrett, P. 1995, in ASP Conf. Ser. 77, ADASS IV, ed. R. A. Shaw, H. E. Payne, & J. J. E. Hayes (San Francisco: ASP), 492

Kowalkowski, J., Piccoli, L., & Simone, J. 2009 Workflow Execution Framework: Models, Provenance, and Fault-Tolerance, http://cd-docdb.fnal.gov/cgi-bin/ShowDocument?docid=2988

Kunszt, P. Z., Szalay, A. S., Csabai, I., & Thakar, A. R. 2000, in ASP Conf. Ser. 216, ADASS IX, ed. N. Manset, C. Veillet, & D. Crabtree (San Francisco: ASP), 141

Tucker, D. L., et al. 2006, Astronomische Nachrichten, 327, 821

Thakar, A. R., Szalay, A., Fekete, G., & Gray, J. 2008, Computing in Science and Engineering, 10, 30

[11] `http://portals.omg.org/dds`