FERMILAB-CONF-11-008-E

A prototype for JDEM science data processing

Erik E Gottschalk

Fermi National Accelerator Laboratory, Batavia, IL 60510, USA
http://www.fnal.gov/

E-mail: erik@fnal.gov

Abstract. Fermilab is developing a prototype science data processing and data quality monitoring system for dark energy science. The purpose of the prototype is to demonstrate distributed data processing capabilities for astrophysics applications, and to evaluate candidate technologies for trade-off studies. We present the architecture and technical aspects of the prototype, including an open source scientific execution and application development framework, distributed data processing, and publish/subscribe message passing for quality control.

1. Introduction

Fermilab is developing a prototype for science data processing and data quality monitoring for the Joint Dark Energy Mission (JDEM). JDEM was being developed jointly by the National Aeronautics and Space Administration (NASA) and the U.S. Department of Energy (DOE) to build a space telescope to study dark energy using multiple, complementary techniques. These included the study of baryon acoustic oscillations, Type Ia supernovae, and weak gravitational lensing. Several months ago the Astro2010 Decadal Survey [1] recommended the Wide Field Infrared Survey Telescope (WFIRST) as the highest priority large-scale space mission for the coming decade. WFIRST is based on a JDEM design (JDEM Omega concept) and includes dark energy research and the study of exoplanets as science goals. Although the future of WFIRST is unclear at this time, we anticipate that much of the work done for JDEM will be incorporated into the WFIRST design. In this regard, the design of a science data processing system for JDEM is relevant to WFIRST as well.

The JDEM design consists of a space telescope and ground systems. The ground systems include the Ground Data System (GDS), which is responsible for operations, receiving and monitoring telemetry, and several levels of science data processing. The first level of data processing is performed by the Science Operations Center (SOC), a key component of the GDS. The SOC is responsible for receiving science data, removing instrument signatures from the data, and applying calibrations. These tasks and subsequent data reduction tasks are the primary motivation for the development of our prototype.

The prototype is referred to as the JDEM Demonstration Data Processing System (JDDPS). JDDPS is designed for scientists and engineers developing simulations as well as data processing and quality monitoring software. The significance of the JDDPS development effort is that existing science data processing systems are inadequate as a basis for new systems, and are not easily applied to missions for which they were not specifically designed. Without development efforts, such as ours, aimed at developing a more generic approach to science data processing, future missions are likely to continue developing their own custom solutions.

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

2. Science driver

JDDPS is designed to provide an execution framework for simulations and science data processing. The science driver for JDDPS is the study of baryon acoustic oscillations (BAO), which entails the determination of length scales of acoustic waves that propagated in the early universe. Our approach to BAO science is based on slitless spectroscopy (for example, see [2]). This involves simulations of slitless spectra for mission concept studies and subsequent processing of the simulated data to assess the ability to achieve BAO science goals.

The JDEM BAO experiment analyzes two types of data: images and slitless spectra. The images are used to identify and measure the positions of stars and galaxies. The slitless spectra (see Figures 1 thru 4) are used to obtain emission-line redshifts (using the H-alpha emission line) of galaxies, which are the primary dataset for the BAO analysis. Slitless spectra of stars are used as part of the calibration process.

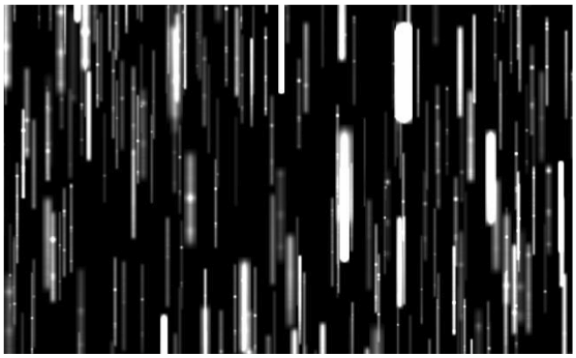

Figure 1. Slitless spectra generated for one telescope *roll angle* (described in the text). The brightest spot on a spectrum usually corresponds to the H-alpha emission line.

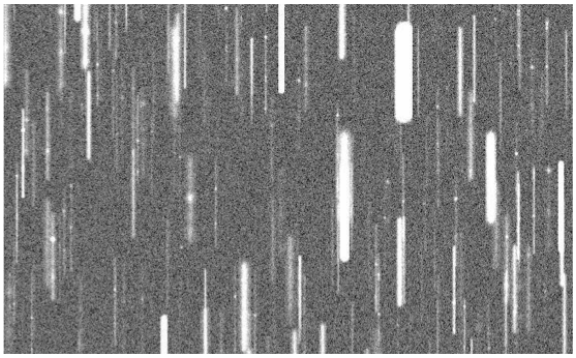

Figure 2. Slitless spectroscopy image generated for the same field shown in Figure 1 with photon noise and random noise included in the simulation.

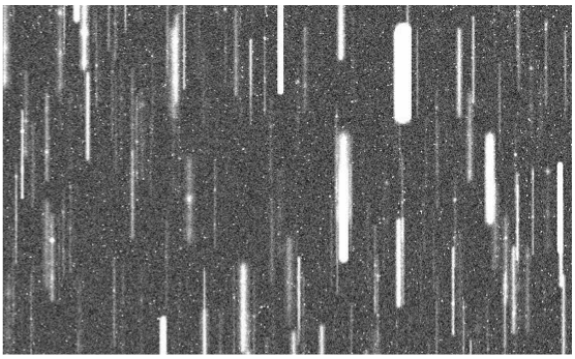

Figure 3. Slitless spectroscopy image of the same field shown in Figure 1 with photon noise, random noise, and cosmic rays included in the simulation.

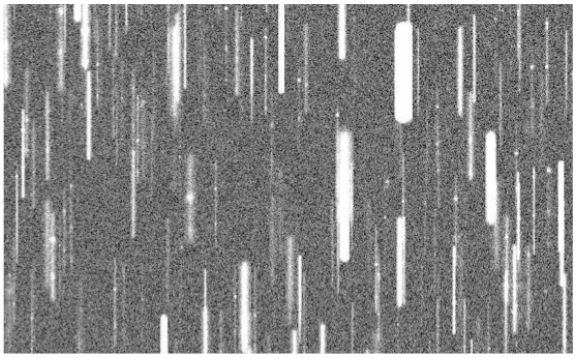

Figure 4. The resulting slitless spectroscopy image after applying an up-the-ramp sampling algorithm to remove cosmic rays from the image shown in Figure 3.

Our simulations thus far have focused exclusively on slitless spectra, since the ability to extract and measure slitless spectra is crucial for the design of space-based dark energy missions. The simulations use input catalogs that provide a realistic representation of the sky, in particular with respect to the number density, sizes, flux distributions, and emission-line intensities of

galaxies and with respect to the number density and flux distributions of stars. At this stage in the development of our simulations the imaging portion of the simulations is presumed to have run successfully, creating a list of candidate objects for input to spectroscopic processing. The spectroscopic portion of the simulation uses the input catalogs to create realistic slitless spectroscopic images, referred to as *spectroscopic frames*. Input parameters include telescope, instrument, and detector properties and knowledge of the zodiacal light background. At present the instrument consists of a single detector, and the telescope orientation (referred to as the *roll angle*) may be rotated about the center of the detector.

Spectroscopic frames and the list of candidate objects are input to the spectroscopic data analysis workflow (see Figures 5 and 6). The workflow consists of four stages of data processing that include optimal spectroscopic extraction, spectral line detection, position fitting for stars, and combining four different roll angles to determine redshifts. The workflow is currently running on a development system at Fermilab.

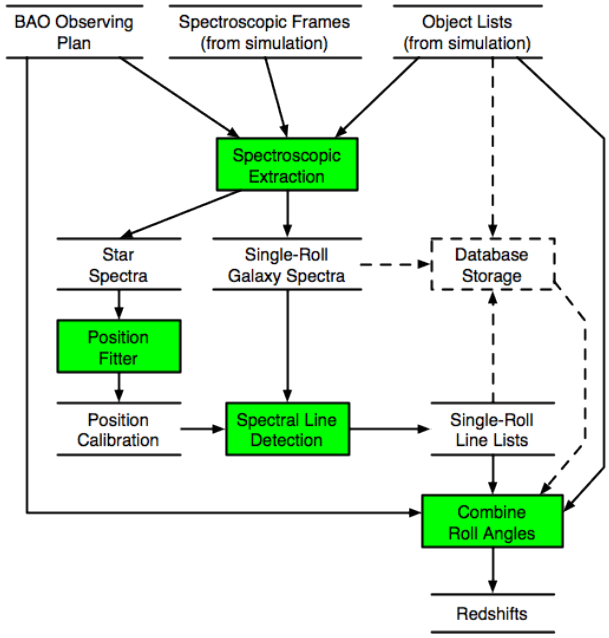

Figure 5. Workflow diagram for slitless spectroscopy showing data processing stages (identified as green-shaded rectangles and described in the text) that take simulated input data, create intermediate data products, and produce redshift measurements for slitless spectra. Solid lines with arrows indicate input and output data flow for each data processing stage in the analysis, and dashed lines with arrows indicate data products that are written to database storage (dashed rectangle) and then retrieved to complete the analysis.

3. Requirements and architecture

JDDPS is designed to operate workflows in a distributed computing environment with data processing stages distributed across multiple processing nodes. This is unlike data processing in high-energy physics, where different workflow stages are usually combined in a single framework executable and deployed on many processing nodes to process data in parallel. For JDEM we have adopted the approach that is more common in astronomy and astrophysics, where a workflow (also referred to as a *pipeline*) typically consists of multiple executables running on different nodes, each performing a different part of the data processing.

The design of JDDPS is based on forty stakeholder requirements that address the needs of scientists. The scientists include software developers and users of JDDPS, and stakeholder requirements are categorized according to different roles. We define four roles: scientist algorithm developer, scientist workflow developer, scientist data analyst, and operator. Three of the roles are usually (but not exclusively) associated with work done by scientists, so we use the term "scientist" to define these roles. The "operator" role is often performed by a scientist who operates data processing workflows, but we do not view this role as requiring much of a

background in science. Other roles associated with software development are defined elsewhere (for example, a role for software infrastructure developers is defined in the JDDPS system requirements).

The JDDPS architecture is designed to provide services (for example, file and database access services) as well as workflow design, planning and scheduling tools. Figure 6 shows the slitless spectroscopy workflow described previously, and identifies the services and tools provided by JDDPS. There are three types of services shown in the figure. *Application services* include application programming interfaces (APIs) for access to FITS files (a standard file format used in astronomy and astrophysics), catalog queries, and quality control. *Data services* provide access to input data for workflows, workflow parameters, databases for output data written to catalogs, and quality control (QC) data. For example, in our prototype the *image store* provides access to images and slitless spectra that will be processed by the slitless spectroscopy workflow, and redshift measurements are written to an output *catalog* for objects found in the images. *Execution services* provide support for control and monitoring of data processing systems, and provide the ability to submit data processing campaigns. We use the term "campaign" to refer to one or more workflows that process data to achieve a particular goal. For example, a campaign can consist of multiple data processing jobs needed to process all slitless spectra acquired during a 24-hour period of data acquisition.

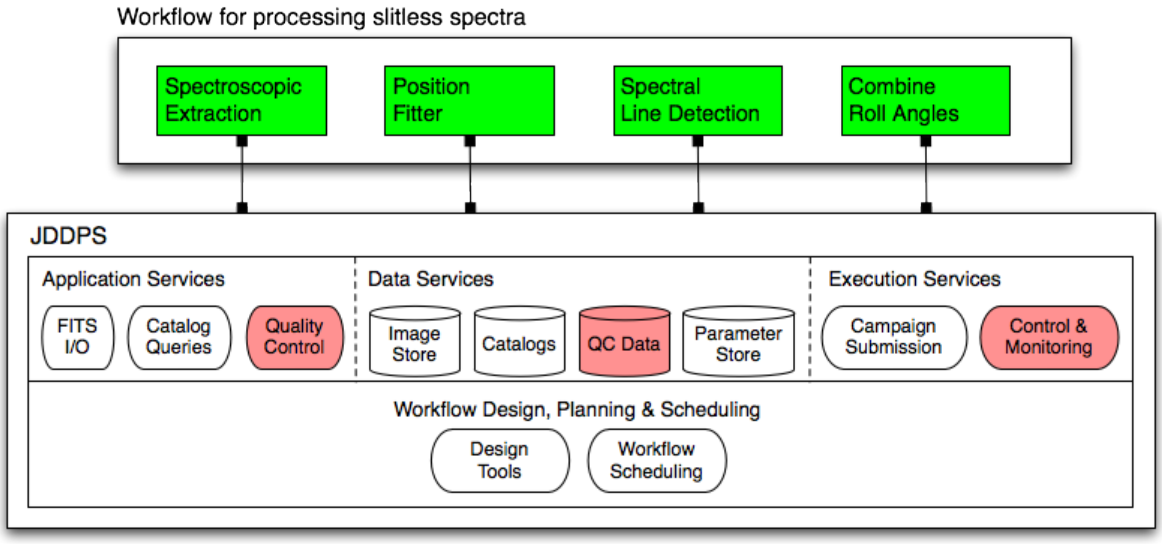

Figure 6. JDDPS architecture diagram showing slitless spectroscopy workflow stages (identified as green-shaded rectangles and described in the text) and the services and tools provided by JDDPS. Services that provide support for quality control are highlighted in red.

4. Software technologies

To implement JDDPS we have chosen several technologies, including an open source scientific execution and application development framework, an open source publish/subscribe message passing system, and relational database software. The chosen technologies are evaluated in the context of the JDDPS prototype and may change over time. We selected Kepler [3] as the scientific execution and development framework after evaluating three generic workflow systems (Swift [4], Pegasus [5], and Askalon [6]) for science data processing in a distributed computing environment. All three workflow systems failed to satisfy important requirements pertaining to workflow specifications, provenance tracking, progress tracking, and ease of use. Since these

workflow systems are under active development, new versions may need to be re-evaluated in the future. We are now evaluating Kepler to determine if it satisfies the requirements.

For publish/subscribe message passing we selected the Data Distribution Service (DDS) publish/subscribe standard from the Object Management Group. The main function of DDS in our prototype is to provide message passing for quality control (QC) data in a distributed computing environment. The overall goal is to reduce the amount of time spent on routine tasks associated with QC, and to reduce the number of ways of interacting with the QC system. For JDDPS we are evaluating the OpenSplice [7] implementation of the DDS standard.

5. Deployment model

Figure 7 shows the JDDPS deployment model with JDDPS functions, their relationships, and the communication domains in which the functions are performed. The cloud shape in the figure denotes a *computation domain* that represents a computing cluster with its own network. Each pill shape denotes a function performed in the system, and each circle denotes a user function. Rectangles are used to denote message-passing infrastructure, and cylinders denote data storage systems. The figure shows which functions are performed within the Fermilab network, and which functions may also take place on a wide-area network.

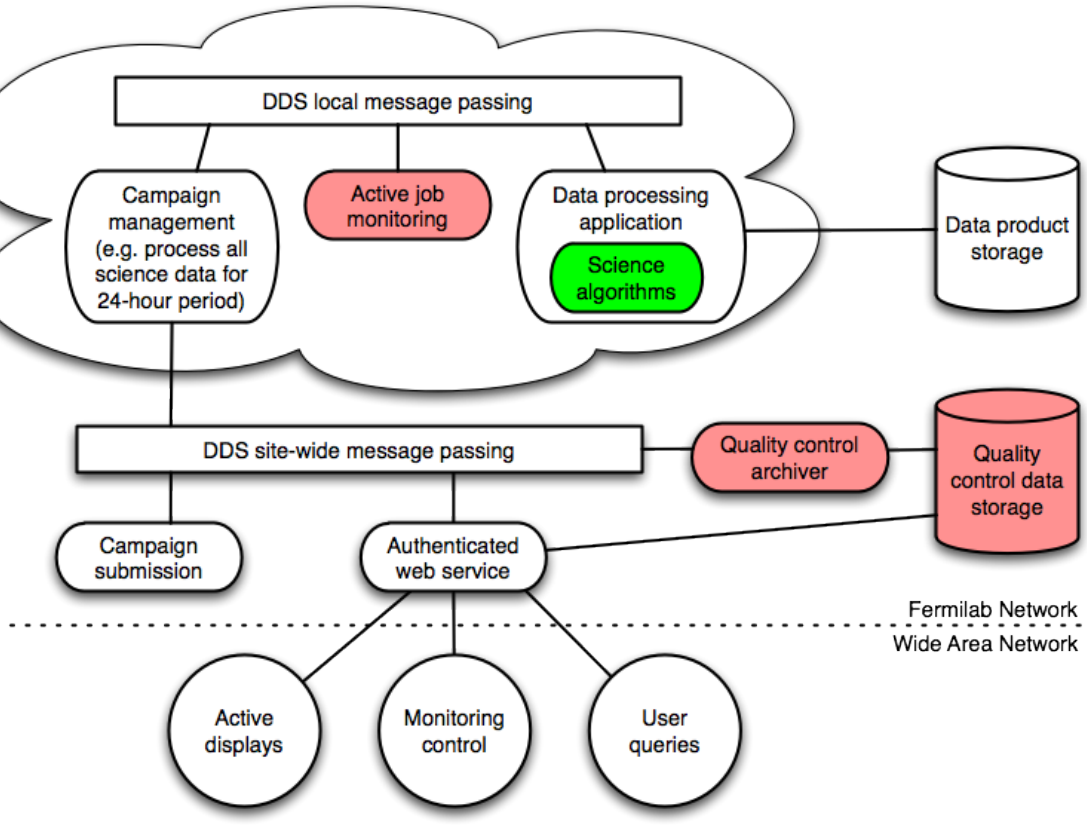

Figure 7. JDDPS deployment model showing JDDPS and user functions, message-passing infrastructure, and data storage systems. The cloud shape denotes a computation domain (described in the text). Science algorithms (shaded green) operate in a data processing application, and quality control functions and storage are highlighted in red.

The *computation domain* represents a single deployment of a workflow performing a task. A workflow consists of a sequence of operations needed to run a JDDPS campaign. A production

system may be comprised of one or more active computation domains, each performing different tasks. The computation domain includes the functions shown in Figure 7, linked by DDS local message passing. The main function is the *data processing* function, which consists of an application with one or more science algorithms (as shown in Figure 6). Individual JDDPS jobs are monitored by the *active job monitoring* function, and the *campaign management* function is responsible for managing all of the jobs needed to process data for an entire campaign.

The campaign management function communicates with functions outside the computation domain through a DDS site-wide message passing system, which operates within the Fermilab network. The figure shows a *campaign submission* function that is used to coordinate the tasks associated with computation domains. Data from a computation domain are transported to a *quality control archiver* that stores data in a data storage system. Data access is provided through an *authenticated web service* that gives users access to all supported user functions. Figure 7 shows user functions, which are available inside and outside the Fermilab network, at the bottom of the figure. These include *active displays* for monitoring QC data, a *monitoring control* function that gives users control of the system, and *user queries* that provide access to archived QC data.

6. Summary

Fermilab is developing a prototype for science data processing and data quality monitoring for dark energy science. The science driver for the JDDPS prototype is the study of baryon acoustic oscillations, and the approach is based on simulations and data analyses of slitless spectra to assess the ability of a future space-based mission to achieve BAO science goals.

In designing JDDPS we have adopted a more rigorous systems engineering approach to software development compared to what is usually done for high-energy physics computing. We identified stakeholders, determined stakeholder needs, and implemented a formal requirements management and traceability process. We included quality control from the beginning by addressing the need for science data quality monitoring and execution environment monitoring as an essential feature of JDDPS. We included workflow management aspects from the beginning, so that scientists will be able to assemble, configure and run science data processing workflows and campaigns, all within one system.

Acknowledgments

In recognition of the dedicated work of the JDEM team at Fermilab, I wish to thank members of the JDEM SOC team: Stu Fuess, Jim Kowalkowski, Igor Mandrichenko, Eric Neilsen, Marc Paterno, Vince Pavlicek, and Vladimir Podstavkov. A special thanks goes to team member and Deputy Project Scientist for the JDEM DOE Project Office, Stephen Kent. Fermilab is operated by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the United States Department of Energy. Funding is also provided by LBNL under Contract No. DE-AC02-05CH11231.

References

[1] "New Worlds, New Horizons in Astronomy and Astrophysics," Washington, D.C.: National Academies Press, 2010, prepublication available as ISBN-10: 0-309-15796-X.
[2] Pasquali A, Pirzkal N, Larsen S, Walsh J R and Kummel M 2006 *The Publications of the Astronomical Society of the Pacific* **118** 270
[3] https://kepler-project.org/
[4] http://www.ci.uchicago.edu/swi-ft/index.php
[5] http://pegasus.isi.edu/
[6] http://www.askalon.org/
[7] http://www.opensplice.com/