THE NATIONAL IGNITION FACILITY DATA REPOSITORY*

R. Carey, P. Adams, S. Azevedo, R. Bettenhausen, R. Beeler, C. Foxworthy, T. Frazier, M. Hutton, L. Lagin, S. Townsend, Lawrence Livermore National Laboratory, Livermore, California U.S.A.

Abstract

NIF is the world's largest and most energetic laser experimental system, providing a scientific center to study inertial confinement fusion [1,2] and matter at extreme energy densities and pressures. This presentation discusses the design, architecture, and implementation of the NIF Data Repository (NDR), which provides for the capture and long-term digital storage of peta-scale datasets produced by conducting experimental campaigns. The NDR is a federated database that provides for the capture of: experimental campaign plans, machine configuration & calibration data, raw experimental results and the processed results produced by scientific workflows. The NDR provides for metadata, pedigree, quality, effectivity, versioning and access control for each of the data categories. A critical capability of the NDR is its extensive data provisioning capabilities and protocols that enable scientists, local and remote alike, to review the results of analysis produced by the NDR's analysis pipeline or to download datasets for offline analysis. The NDR provides for the capture of these locally-produced analysis results to enable both peer review and follow-on automated analysis.

NDR FEDERATED DATABASE

The NIF federated database architecture transparently integrates multiple autonomous database systems. Through intelligent information system applications and data abstraction, the federated database provides a uniform user view, enabling users and applications to store and retrieve data in multiple autonomous databases/schemas with integrity and consistency. NDR represents the fully-integrated logical composite of all constituent databases and schemas. The constituent databases are standardized on the Oracle information system product suite.

Extract. Transform, Load

Bulk data movement within NDR is done using extract, transform, and load (ETL) processes. A large portion of the experimental result data from NIF are 2-D images. NIF generates approximately 66 Tera-Bytes per year. ETL processes are used to ingest image data and add structured meta data for long term archive. Individual experiments on NIF generate over 20 thousand images for laser alignment, shot time capture, and post shot optics damage inspection. Automated data analysis systems are triggered by ETL completion events and perform just-in-time analyses that guide decision making for subsequent experiment configurations and future optics refurbishment. Recent data is stored on-line and a tiered storage architecture migrates least recently used data to near-line storage. Information lifecycle management (ILM) policies are implemented as data driven aging algorithms and allow having different migration strategies for different types of data.

Federated Transactional Semantics

NDR consists of multiple databases for tracking and maintaining accurate NIF configuration data and experiment results. Movement of NIF parts/components through installation, calibration, and operational qualification requires that various information systems and constituent databases be updated so that NIF experiments are executed and diagnosed accurately, reliably, and safely. Automation of configuration management is required due to NIF's inherent complexity and the significant number of parts (> 6 million).

To automate reliable change management for NIF components, an information system messaging architecture has been defined that provides a general mechanism for updating necessary information systems as a result of NIF service orders. Each type of NIF component that requires change management typically has custom work procedures for installation, calibration, and operational qualification. At certain points in the work procedure, configuration and calibration information must be updated for operational usage.

In addition, this same messaging architecture is used to notify experiment analysis and visualization systems of experiment completion events and availability of raw and processed experiment results. The information system messaging architecture provides for construction and automated delivery of messages between different NIF information systems. The messaging architecture is based upon Oracle Advanced Queuing and messages are defined as XML documents. This architecture provides guaranteed message delivery, priority, retries, message history, scheduling, tracking and event journals, and internet integration.

SCIENTIFIC WORK FLOW

Preparation and analysis of data from a NIF target experiment requires the ability to quickly analyze and interpret all data captured from many different specialized diagnostic instruments at the Target Chamber. The timeliness of this analysis is driven by the fact that the results of one experiment will likely guide the configuration of the next. During the experiment preparation phase, these analyses address issues such as laser alignment, focus, and other tuning parameters necessary to achieve successful fusion ignition. Following a target experiment, the analyses are focused on interpreting the raw data collected from the diagnostic instruments. The results of these analyses describe the time resolved physical phenomena produced by the impacted target. These types of measurements are critical to understanding the science of all NIF missions including stockpile stewardship, high energy density (HED) science, astrophysics, and clean fusion energy.

In recent years, scientific workflow systems[4,5] have emerged as key tools to integrate different computing and data analysis components, and to control the logic between computing tasks. The NIF Shot Data Analysis Engine is an integration of commercial workflow tools and messaging technologies applied to a scientific data analysis application domain. The NIF Shot Data Analysis Engine provides automatic triggering of analysis upon arrival of new data; analysis workflow sequencing; data provisioning (ETL) from various data sources; data mapping to analysis functions written in Interactive Data Language (IDL); and results archiving with pedigree (a record of the data inputs and the specific version of analysis software used).

This scalable, parallel data analysis architecture utilizes a commercial, industry standard Workflow Processor called Business Process Execution Language (BPEL) whose strength focuses on "orchestration" and its ability to integrate/interface to external systems through Web Services technology. The BPEL Workflow Processor takes analysis workflow requests and performs the following three functions: (1) orchestrates the sequence of analysis steps; (2) integrates, transforms, and transfers data between various data repositories and the analysis modules; and (3) schedules the appropriate analysis on a cluster of compute servers and stores the results in the NDR Content Management System.

DATA STEWARDSHIP

The primary focus of data stewardship is determining an organization's data warehouse content, maintaining common definitions, assuring data quality, and managing appropriate access. Data from NIF will be accumulated and maintained for several decades. Below are brief descriptions of NDR standards and practices selected for data capture, data provenance, data identification, and content management. These collectively address core data stewardship values of confidentiality, integrity, and availability.

Data Capture

Raw data from NIF comes in several forms. Laser diagnostic data consists of scalar, vector, and beam imagery. These data elements have well defined structure and are managed in a relational archive database. The archive of laser beam images is described by a complete set of metadata and the images themselves are stored in a blob (binary large object) column. For target diagnostics, a standard plug-and-play architecture and communications protocol have been defined to allow collaborating scientists to design and build diagnostic instruments that can be "plugged into" the NIF control system. The protocol specifies a command set for registering, activating, timing, and data capture for target diagnostic instruments. In addition to the communications protocol, NIF has standardized on the open source HDF5 (Hierarchical Data Format) self describing file format for calibration data, recording background data, and shot time diagnostic data capture. HDF5 provides a versatile and portable data model that can represent very complex data objects and a wide variety of metadata. The HDF5 software library runs on a range of computational platforms and implements a high-level API for several programming languages including C++ and Java. This portable architecture for target diagnostics enables NIF to support external scientific collaboration and experimentation.

Data Provenance

One of the basic principles of the scientific method is integrity and reproducibility of experimental results. This requires maintaining and publishing a strict chain-of-custody that follows the raw data through to the results calculated with analysis algorithms. Provenance[6] (also referred to as lineage and pedigree) is an acute issue in scientific databases and is central to the validation of data. The NDR supports tracing and recording the origins of data as it becomes refined and transformed into summary information. Derived data must be associated with the specific version of transformations/algorithms applied.

Input data such as instrument calibration, and analysis algorithm revisions, must accompany analysis results and serves to identify pedigree. If inputs are updated or refined, results must be versioned and re-calculated. Each version of the analysis results has its own pedigree. Calibrations must also reflect effective dates (effectivity). Changes to an instrument design likely affect its calibration and data captured on different dates must be paired with the effective calibration data at the time of the experiment.

Data Identification

NDR has adopted internet information architecture (W3C) to identify each member of the archive. Uniform Resource Names (URN) serve as persistent, location-independent resource identifiers and are designed to be globally unique. URNs have the following form:

urn:nif.llnl.gov:archive:GUID

The GUID (globally unique identifier) is a special type of identifier used in software applications to provide a reference number which is unique in any context. GUIDs are automatically generated for new members. The URN lends itself to Web services allowing data to be published in a standardized vocabulary.

Content Management

A content management system[7] is used to manage document work flow needed to collaboratively create, edit, review, index, search, publish, and archive NIF data. The NDR is comprised of data in relational database tables and collections of files. The content management system integrates these data representations into a uniform view that appears as a navigable file system to the user. This architecture supports unstructured and structured data elements. It has the ease of use of a file system and lends itself to presentation via internet applications. The NDR content management system provides access from many client platforms and supports protocol servers for access from Web clients and email clients.

Content management provides a versioning model as part of data provenance. Versions of a document form a document family. A family can be manipulated as a single entity with a version series. The version series can be used to support different versioning schemes.

Content management supports the access control requirements of the NDR. Access control lists (ACL) can be applied to all objects in the repository including directories, files, and database tables. This allows individual scientists to protect intellectual property and share information with a restricted set of colleagues. Objects can be categorized and annotated with properties. Relationships can be defined between selected objects. Groups and/or roles can be defined to facilitate content access.

HED SCIENCE PORTAL

The vision for the NDR is to provide a secure internet facing portal that fosters collaboration among research institutions and universities to enhance the vitality and availability of high-energy-density scientific research.

Existing scientific information Web sites and e-science projects are being studied as examples. The HED portal must support large data sets, be secure and scalable. Researchers need to be able to share tools, algorithms, and data sets. It must accommodate a "Review and Release" policy that requires independent review of materials to be made publically available. Thus, electronic Review and Release protocols need to be established to insure quality, data provenance, and adherence to the scientific method.

The HED science portal will attract researchers from research laboratories and universities around the globe. It will make science of global interest publically available and help establish the National Ignition Facility as a world class experimental science laboratory.

DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

*AUSPICES STATEMENT

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

REFERENCES

[1] E. Moses, "The National Ignition Facility: Path to Ignition in the Laboratory," Fourth International Conference on Fusion Sciences and Applications," Biarritz, France, September 2005.

[2] The National Ignition Facility Web site, https://lasers.llnl.gov,

[3] Azevedo, S., et al, "Automated Experimental Data Analysis at the National Ignition Facility ", ICALEPCS '09 Conference Proceedings, Kobe, Japan, 2009.

[4] Complete Guide to Nuclear Fusion, Fusion Energy and Power Plant Reactor Research, with Encyclopedic Coverage of Facilities and Labs (DVD-ROM), World Spaceflight News, October 2005, CD-ROM: ISBN 1422001199.

[5] Grid Computing: Experiment Management, Tool Integration, and Scientific Workflows, Springer; 1st edition, February 21, 2007.

[6] 2008 BPM & Workflow Handbook - Spotlight on Human-Centric BPM, Future Strategies, Inc.; 1st edition, March 17, 2008.

[7] Yogesh L. Simmhan et.al., "A Survey of Data Provenance in e-Science", SIGMOD Vol. 34, No.3, Sept 2005.

[8] Simon Azriel, "Oracle Content Management SDK: Concepts and Architecture", Oracle Corporation, August 2005.