Data Systems Modernization (DSM) Project: Development, Deployment, and Direction

Robert M. Whitten Jr., ORNL

ABSTRACT: The Data Systems Modernization (DSM) project was undertaken to consolidate and update the current information systems of the Oak Ridge Leadership Computing Facility (OLCF). The project combined the Resource Allocation and Tracking System (RATS), the New Account Creation System (NACS) and open-source process management and business intelligence software to streamline the data processing systems of the OLCF. This paper will discuss the development, deployment and future directions of this ongoing project.

KEYWORDS: XT5, Cray, RATS, Jaguar, Gaea, Kraken, NACS, DSM

1 Introduction[1]

1.1 ORNL Computing Facility

The ORNL computing facility is home to the Department of Energy's (DOE) Oak Ridge Leadership Computing Facility (OLCF). The facility houses some of the world's largest computational resources. The center is also known as the National Center for Computational Sciences (NCCS).

In addition to DOE resources, the facility also houses systems for the National Science Foundation (NSF) and the National Oceanic and Atmospheric Administration (NOAA). (see the below image)
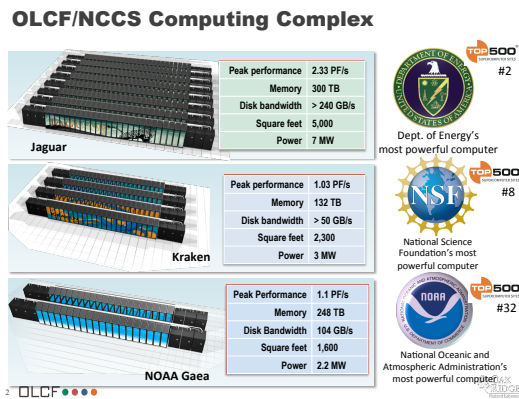
OLCF/NCCS Computing Complex

1.2 What is DSM?

The Data Systems Modernization (DSM) project is a software project that strives to consolidate the various data sinks that exist the NCCS.

DSM is a business intelligence tool as well as a data-warehousing tool. Its architecture can be thought of as a silo-architecture of middle-ware applications. DSM has the primary function of acting as an extract-transform-load (ETL) tool.

DSM is a combination of several middle-ware tools that are in use at the NCCS. These tools include the Resource Allocation and Tracking System (RATS) the New Account Creation System (NACS), the DowntimeDB, and HPSS stats. Components of these middle-ware tools use a combination of technologies such as MySQL, LDAP, and scripts to access data. DSM adds additional components such as ProcessMaker, LDAP synchronization scripts, and LogiXML.

2.0 Middle-ware Applications

2.1 RATS

The Resource Allocation and Tracking System (RATS) is the primary data source of the NCCS. The system tracks allocation usage on a per-user basis. User allocation usage is then charged on a per-project basis.
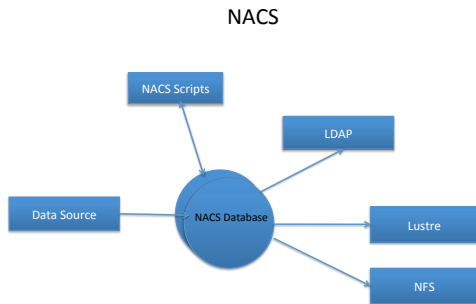
The below diagram describes the components of RATS.
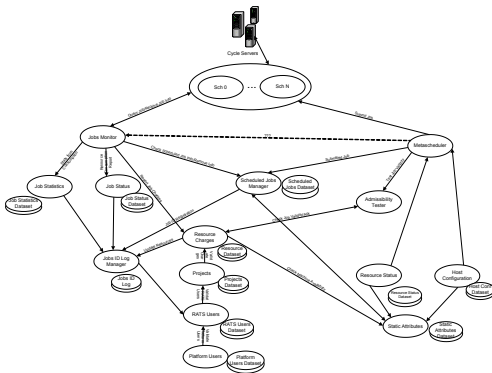
2.1 NACS

The New Account Creation System (NACS) is a 'last mile' system designed to take user applications from a web-based interface and create system user accounts and file system spaces.

NACS uses a push-pull architecture to update necessary system resources such as LDAP and file systems.

NACS

2.2 DowntimeDB

The DowntimeDB provides a mechanism for system administrators to manually enter downtime information. The data is then used to provide management reports.

Data Source → Downtime Database → Reports
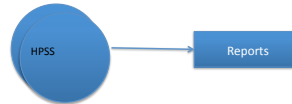
8 OLCF ●●●●

2.3 HPSS Stats

Users store data on the High Performance Storage System (HPSS). The HPSS Stats component reads the appropriate metadata stores to determine the amount of data stored for archival.

HPSS Stats

• Data read directly from HPSS metadata

HPSS → Reports

9 OLCF ●●●●

3.0 Why DSM?

The issue with running multiple middle-ware applications is data redundancy and inconsistency. Having user data common among the different data sinks causes inaccurate data to exist in one area. This problem is exacerbated by not having a consistent interface to entering updates or changes to the various data sinks.

DSM's primary goal is to provide a consistent interface to the data stored for NCCS users and resources. A secondary goal of DSM is provide substantially better reporting mechanism to allow management personnel an easier time at generating needed reports.

2

4.0 Additional DSM Components

4.1 ProcessMaker

ProcessMaker is open source workflow software solution. It provides a way for users to translate their business processes into an automated system capable of generating emails and notifications to personnel of pending workflow items.

In the initial phases, DSM will use ProcessMaker for account and project application creation. In later phases, DSM will migrate increased functionality to ProcessMaker.

4.2 Interface Scripts

DSM provides scripts developed at ORNL to allow staff to modify user, group, project, and allocation attributes. These scripts are mostly written in Python. The functionality provided by these scripts will eventually be migrated to ProcessMaker.

4.3 LogiXML

LogiXML is a business intelligence (BI) tool that provides enhanced ability to produce reports and information based on data stored in DSM. It is a commercial product similar to Crystal Reports or any of the other BI products on the market.

5.0 DSM Deployment Schedule

5.1 Phase 1

In this phase, DSM will be deployed on NOAA systems. This instance of DSM will consist of the RATS and NACS components. The difference will be in how data is entered into the system. A synchronization script will retrieve data from a remote LDAP instance and store directly into a local LDAP instance.

There will be neither LogiXML nor ProcessMaker component. NOAA is a remote user facility and provides their own application and reporting functionality.

This phase was completed in the first quarter of fiscal year 2011.

5.2 Phase 2

In this phase, DSM will be deployed on DOE systems. LogiXML and ProcessMaker components will be deployed. Additional access scripts will be created to aid personnel in accessing DSM.

This phase is targeted for completion in the fourth quarter of fiscal year 2011.

5.3 Phase 3

In this phase, additional requirements that were gathered during previous phases will be considered. Added functionality beyond account creation will be implemented in ProcessMaker.

RATS has an open source descendent, which has recently (as of April 2011) been released. Knows as DataMux (available on Source Forge), this release has enhanced components that may replace the core components of RATS as in deployed with DSM.

Consolidation of the NOAA and DOE instances of DSM will also be considered.

6.0 Conclusion

DSM has to goal of consolidating and augmenting the software data systems of the NCCS. By combining feature of RATS, NACS, and other systems, DSM will become the authoritative data source of the NCCS.

About the author

Robert Whitten Jr. is a member of the User Assistance and Outreach Group at Oak Ridge National Laboratory and can be reached at whittenrm1@ornl.gov.