OODTProposal

OODT, a grid middleware framework for science data processing, information integration, and retrieval.

Abstract

OODT is a grid middleware framework used on a number of successful projects at NASA's Jet Propulsion Laboratory/California Institute of Technology, and many other research institutions and universities, specifically those part of the:

- National Cancer Institute's (NCI's) Early Detection Research Network (EDRN) project - over 40+ institutions all performing research into discovering biomarkers which are early indicators of disease.
- NASA's Planetary Data System (PDS) - NASA's planetary data archive, a repository and registry for all planetary data collected over the past 30+ years.
- various Earth Science data processing missions, including Seawinds/QuickSCAT, the Orbiting Carbon Observatory, the NPP Sounder PEATE project, and the Soil Moisture Active Passive (SMAP) mission.

From the OODT website:

It's middleware for metadata:

- Transparent access to distributed resources
- Data discovery and query optimization
- Distributed processing and virtual archives

It's a software architecture:

- Models for information representation
- Solutions to knowledge capture problems
- Unification of technology, data, and metadata

Proposal

OODT is an established open source project, with 9+ years of existence, and deployment at universities, federal research institutions, other NASA centers, and the NIH (it won runner-up NASA software of the year in 2003). It has a strong community of those that operate and support its growth. Our proposal is to bring OODT into Apache to strengthen its support and its capabilities even further on the laurels of Apache's brand and its growing huge community of developers from all over the world. In short, bringing OODT into Apache will significantly enhance OODT's widespread use, will likely improve its codebase, and furthermore will help Apache philosophy and community spread into OODT's already large community-base reaching across government, academia and industry.

OODT will be, to the best of our knowledge, the *first* grid community project to bear the Apache brand. By *grid* technology, we mean a technology that provides the ability to create *virtual organizations*, as originally described by Kesselman and Foster in their seminal paper on grid computing. OODT provides both computational *and* data grid support, and is built with a component-philosophy. OODT includes components that allow for virtual information integration across organizations (provided by the *Profile*, *Product* and *Query* server components), and that allow for distributed data management and processing across heterogeneous virtual organizations (provided by the Catalog and Archive Service set of components, including *File Manager*, *Workflow Manager* and *Resource Manager*).

Each set of components exist as independently organized Maven2 projects, that reference each other (where appropriate), forming a layered set of components and a framework for grid computing.

Background

OODT is an established project within JPL and in use at several NASA centers, as well as univerities, and other government organizations and industrial collaborations. Chris Mattmann, a JPL employee, and ASF PMC (Lucene) and Committer (Nutch, Tika), has been working for the past 2 years on obtaining the necessary permission from JPL to release OODT into Apache. After initially being stalled, JPL has granted permission to allow OODT into Apache.

Through his academic relationship with Justin Erekrantz, Apache President, and through their collective Ph.D. studies, OODT has been discussed between Chris and Justin on several occasions, and Justin offered to help champion OODT into the Apache Incubator when JPL was ready to release OODT. In December 2009, that permission was granted.

This proposal is the result of the above efforts and related discussions. Some alternatives to incubation, like Apache Labs came up during the discussions but we believe that taking the project to the Incubator is the best way to start growing a viable Apache-based community to sustain OODT. Furthermore, given its larger code base and existing sub-projects, the goal would be for OODT to leverage the incubator to graduate into Apache's first top-level grid project, rather than graduate into a sub-project of an existing TLP.

Rationale

Grid computing has been around for the past 10 years and has gained widespread notoriety and attention in industry and academia. Scientific collaborations are increasingly virtual and require the capabilities (data and compute) of thousands of computers and resources that span organizations. There are a number of existing grid technologies (Globus being the most popular, DSpace, iRODS/Storage Resource Broker, see this paper for a full study), however Apache has **no current grid technology** under its umbrella and world reknown think tank. Morever, efforts are few and far between in terms of standing up Apache-based software that is applicable to the scientific community and grid community outside of use of fine-grained components in these systems. Other open source organizations (e.g., the Global Organization for Earth System Science Portals, GO-ESSP) have embraced the construction of such technology and there is a lot of work going on, e.g., at NOAA. This proposal aims to remedy this fact and to bring scientific data management/grid software into the Apache family and its worldwide community.

OODT is a widely successful grid project with applicability and existing deployments across broad-reaching domains (planetary and earth sciences, cancer research/biomedicine, climate modeling and atmospheric science, etc.). The marriage of OODT and Apache will engender OODT's widespread, global use via the Apache brand, and will make Apache a player in the grid/scientific data community.

Initial Goals

The initial goals of the proposed project are:

- Stand up a sustaining Apache-based community around the OODT codebase.
- Active relationships and possible cooperation with related projects and communities.
- Refactor and bring up-to-date the OODT profile and product server components.
- Explore various underlying communication substrates. OODT currently uses REST (via its Web-Grid component).
- Create configuration-based OODT deployments. Currently the deployments are primarily code-based, or the configuration is strewn about the various sub-components. The goal would be to bring this configuration under a single umbrella project. The idea would be to create science data pipelines from configuration.
- Explore Python-based client and server implementations of OODT and implementations in other languages (Ruby).

Current Status

Meritocracy

Many of the proposed initial committers are familiar with the meritocracy principles of Apache, and have already worked on the various source codebases (contributing via patches, emails, JIRA issues, and in Mattmann's case, as a Nutch, and Tika committer, and Lucene PMC member). We will follow the normal meritocracy rules also with other potential contributors.

Community

There is an existing, established community of developers and users of OODT within over 40 centers at NASA, NIH, DOE and academia, however there is no Apache OODT community as of yet. Our principal goal of this effort is to leverage the Apache Incubator to grow an Apache community base (in addition to OODT's existing community), and to build a self-sustaining community around this shared vision, and eventual Apache TLP status for OODT. With many sub projects (CAS, Product/Profile servers, Query Server, Web-grid, commons, etc.), OODT should attract a broad audience of developers with various interests.

Core Developers

The initial set of developers comes from NASA JPL, and with various backgrounds, with different but compatible needs for the proposed project. JPL is home to data management and grid projects spanning the domains of cancer research/bioinformatics, earth science, planetary science, astrophysics, and climate modeling.

Alignment

As Apache's first grid-based framework will likely be widely used by various open source, scientific and commercial projects both together with and independent of other Apache tools. With OODT's existing community we will also bring developers and organizations outside of Apache into the Apache ecosystem.

Known Risks

Orphaned products

OODT has supported itself through successful deployments at NASA, at the U.S. National Institutes of Health (NIH), and recently at DOE-based laboratories and at academic centers. Further, OODT has been an active participant in IEEE/ACM-based conferences and meetings/journal publications over the past 9 years. There is active support on several existing NASA earth science missions, and the team at JPL is experienced and will continue to champion and develop OODT in the Apache area.

Our goal is to take OODT from the early stage of Apache Incubation into a thriving Apache top-level project, and leverage it in the existing manner at NASA, the NIH, at DOE, and in academia and industry. Since OODT is a grid framework, it depends on many external services and projects, no one of which controls OODT's code-base.

We feel that the time is ripe to bring OODT into Apache and to grow the community of developers who maintain OODT. We feel that Incubation will bring a slew of industry-based developers (and even those in academia, and government) who have no prior experience with OODT, but who could use OODT at their jobs and who are attracted to the brand name and community that Apache brings. We want to attract such developers to become part of the core OODT development team, and project management aspect.

Inexperience with Open Source

All the initial developers have worked on open source before and at least one (Mattmann) is a committer and PMC members in the Apache Lucene ecosystem. Sean Kelly is a well-respected Plone committer and has made several open source contributions over the years to FreeBSD and other software. Foster, McCleese and Woollard have all contributed to Apache projects by way of email, mailing lists, issue reporting and testing.

Homogenous Developers

The initial developers come from a variety of backgrounds and with a variety of needs for the proposed framework.

Reliance on Salaried Developers

All of the proposed initial developers are paid to work on this or related projects, but the proposed project is not the primary task for anyone.

Relationships with Other Apache Products

OODT is related to at least the following Apache projects. None of the projects is a direct competitor for OODT, but there are many cases of potential overlap in functionality.

- Apache Lucene - The family of Lucene products that implement search services are naturally of use in a grid environment such as OODT. In fact, OODT has integrated with many of these projects (Tika, SOLR and Lucene-java) already. We see OODT as a grid environment that makes use of search services.
- Apache UIMA - The UIMA project provides a framework and pluggable tools for analyzing text content and extracting information. Example tools include language identification, sentence boundary detection and "entity extraction" - finding references to people, places and organizations. OODT is related to UIMA in the sense that it is a framework to provide pluggable connections to content and information, but the focus of OODT is on scientific data sets, and additional on repositories and catalogs/registries that catalog information about those datasets and that store the physical bits. Further, OODT is a grid technology, meant to enable the creation of virtual organizations, which is not UIMA's focus.Finally, OODT contains both an information integration component, as well as a science data processing component, which UIMA does not.

OODT is also related to Apache projects involving databases, such as the Apache DB project, however scientific data is not limited to traditional DBMS'es and involves both structured and un-structured information. However, there is likely much leveraging that can occur as OODT can be updated to remove Hibernate-like dependencies, and replace them with Derby-like dependencies.

A Excessive Fascination with the Apache Brand

All of us are familiar with Apache and have a respect for its brand and community. Though all of the proposed committers besides Mattmann have not participated in Apache projects as committers, and PMC members, many of them (McCleese, Foster, Woollard, Kelly) have contributed via issue comments, patches, and tests for Apache projects (including Maven, Tika, SOLR, and Lucene). Furthermore, some of the proposed committers (Kelly) are major contributors in other open source communities (e.g., Plone and Python). We feel that the Apache Software Foundation is a natural home for a project like this. OODT brings a credible, major grid-based software into the Apache community, and Apache brings a huge community of eager and world-class developers to help grow OODT's strengths and applicability across projects and domains.

Documentation

There is a wealth of documentation available on OODT. The best starting point is the existing OODT JPL website (which will be ported to be sync'ed or just a pointer to the Apache website)http://oodt.jpl.nasa.gov

- OODT website at JPL
- Mattmann's OODT paper that appeared at the 28th International Conference on Software Engineering in Shanghai, China.
- Crichton's seminal OODT paper appearing at the CODATA conference at the U.S. National Academies of Science in 2000.
- Google Scholar search on OODT.

Standards and conventions related to OODT include the Dublin Core metadata set, ISO/IEC 11179, the HTTP 1.1 RFC, Grid-based standards including the Open Grid Services Architecture (OGSA), and standards for science data formats including Heirarchical Data Format (HDF), netCDF and OPeNDAP.

Initial Source

OODT will start with seed code donated by NASA JPL via Mattmann and the rest of the initial committers.

Source and Intellectual Property Submission Plan

All seed code and other contributions will be handled through the normal Apache contribution process. Mattmann has been authorized by NASA JPL to lead the contribution of OODT into the Incubator via his existing Apache CLA.

We will also contact other related efforts for possible cooperation and contributions.

External Dependencies

OODT depends on a number of external connector libraries with various licensing conditions. An initial list of such dependencies (taken from one of the OODT sub-components, the CAS file manager) is shown below.

Library	License
commons-codec	AL v2
commons-dbcp	AL v2
commons-httpclient	AL v2
commons-io	AL v2
commons-pool	AL v2
cas-metadata	(to be AL v2)
edm-commons	(to be AL v2)
hsqldb	LGPL v2.1
jug-asl	AL v2
lucene-core	AL v2
xmlrpc	AL v2

There are also some LGPL components that would be useful. Whether and how such dependencies could be handled will be discussed during incubation. No such dependencies will be added to the project before the legal implications have been cleared. Existing LGPL dependencies, such as hsqldb above for the CAS file manager, will be removed and a suitable ASL friendly alternative will be investigated and used to replace the LGPL dependencies.

Cryptography

OODT itself will not use cryptography, but it is possible that some of the external product or profile server or CAS libraries will include cryptographic code to handle features present in various science data formats. The current OODT code base relies on Apache Tika which contains an export control statement regarding cryptographic code per Apache policy. We will follow a similar approach with OODT. Mattmann led this effort in Apache Nutch and saw Jukka Zitting lead this effort in Apache Tika, so he is familiar with this process.

Required Resources

Mailing lists

- oodt-dev@incubator.apache.org
- oodt-commits@incubator.apache.org
- oodt-private@incubator.apache.org

Subversion Directory

- https://svn.apache.org/repos/asf/incubator/oodt

Issue Tracking

- JIRA OODT (OODT)

Other Resources

- OODT Wiki http://cwiki.apache.org/OODT

Initial Committers

Name	Email	Affiliation	CLA
Chris A. Mattmann	mattmann at apache dot org	NASA Jet Propulsion Laboratory	yes
Daniel J. Crichton	crichton at jpl dot nasa dot gov	NASA Jet Propulsion Laboratory	yes
Paul Ramirez	pramirez at jpl dot nasa dot gov	NASA Jet Propulsion Laboratory	yes
Sean Kelly	kelly at jpl dot nasa dot gov	NASA Jet Propulsion Laboratory	yes
Sean Hardman	shardman at jpl dot nasa dot gov	NASA Jet Propulsion Laboratory	yes
Andrew F. Hart	ahart at jpl dot nasa dot gov	NASA Jet Propulsion Laboratory	yes
Joshua Garcia	joshua at jpl dot nasa dot gov	NASA Jet Propulsion Laboratory	yes
David Woollard	woollard at jpl dot nasa dot gov	NASA Jet Propulsion Laboratory	yes
Brian Foster	bfoster at jpl dot nasa dot gov	NASA Jet Propulsion Laboratory	yes
Sean McCleese	smcclees at jpl dot nasa dot gov	NASA Jet Propulsion Laboratory	yes

Sponsors

Champion

- Justin Erenkrantz (jerenkrantz at apache dot org)

Nominated Mentors

- Chris A. Mattmann (mattmann at apache dot org)
- Justin Erenkrantz (jerenkrantz at apache dot org)
- Ross Gardler (rgardler at apache dot org)
- Jean-Frederic Clere (jfclere at apache dot org)
- Ian Holsman (ianh at apache dot org)

Sponsoring Entity

- Apache Incubator