Contribution ID: **112** Type: **Sessions contributions**

Seismology Data Management in VERCE

Monday, 19 May 2014 16:30 (15 minutes)

Seismic data is processed and analysed by researchers across the world. Each researcher deals with terabytes of data (raw and processed) which is hard to manage, process and share adopting exclusively the local facilities. In order to handle this big volume of data, a robust and scalable data management platform is required, allowing storage and discovery of large and heterogeneous datasets. A shared catalog should enable users to search upon standard and community defined metadata, including user annotations and provenance.
Moreover, we aim at minimising the transfer of data among the computational clusters within a distributed processing scenario via the implementation of a federated architecture and standard protocols.
When it comes to sharing of data and resources the privacy of data is a main concern. Each institute requires to retain the complete control over the access permissions on the archived data, establishing and applying data sharing policies in agreement with the research partners.
We will illustrate in this work the approach of the VERCE project to tackle these challenges.

Wider impact and conclusions

This data platform will be used by different services that will be available on VERCE platform. Right now, this is used by VERCE's HPC use case, a forward modelling portal for storing and retrieving data and results which is generated from SuperMUC.
On successful completion of the prototype supporting both forward modelling and cross correlation use case, this setup can be deployed in other partner sites.
We are also considering a global catalog shared by all VERCE users and a query services to fetch the data in a better and simpler way.
Overall this helps seismologists in processing data and do simulations in a better way.

URL(s) for further info

Verce Project
http://verce.eu
Forward modelling portal
http://129.215.213.249:8080/liferay-portal-6.1.0/
iRODS web interface
http://dir-irods.epcc.ed.ac.uk/irodsweb/

Description of work

Our test environment is setup in University of Edinburgh in Opennebula Virtual Machines which is federated with other institutes like CINECA, SCAI, IPGP, ISTerre and INGV
We have an iRODS installation to manage and federate users and sites preserving their policies and users. This iRODS installation provides a set of data management tools for users to upload download and view data and results in iRODS.
The iRODS catalog was relational and do not scale according to our requirement. So we had to use a distributed NoSQL meta-data catalog in MongoDB for data stored in iRODS. To keep the catalog up to date we created microservices to detect changes in data and update the catalog periodically
iRODS can support different file systems, they recently introduced support to HDFS. We have configured a hadoop cluster using virtual machines which can scale easily according to our needs and our current configuration gives user the option to store data in HDFS for archiving and distributed processing.
The data often has to be moved in and out of different HPC resources and the parallel data transfer provided by iRODS was not supported by these resources. With the help of a DSI module from CINECA, we were able to provide a GridFTP Interface for our iRODS installation.

Primary author: Mr MURALEEDHARAN, Visakh (IPGP)

Co-author: SPINUSO, Alessandro (KNMI)

Presenter: Mr MURALEEDHARAN, Visakh (IPGP)

Session Classification: e-Infrastructure Services for Earth Science

Track Classification: Success stories in using e-Infrastructures for research (Track Leaders: E. Katragkou, P. Castejon)