GeoCloud Project Report – GEOSS Clearinghouse

Qunying Huang, Doug Nebert, Chaowei Yang, Kai Liu

2011.12.06

Description of Application

GEOSS clearinghouse is a FGDC, GEO, and NASA project that connects directly to various GEOSS components and services, collects and searches distributively the data and services via interoperable mechanisms. The clearinghouse has the computational challenges of big data (for metadata) and concurrent intensive as well as global machine to machine access.

Operating Organization

Federal Geographic Data Committee (FGDC)

Center of Intelligent Spatial Computing for Water/Energy Science (CISC) of GMU

Community of Interest

Group on Earth Observations (GEO) community

OS and software requirements

Table 1. OS and software requirement for GEOSS Clearinghouse

Specification	Requirements
OS	CentOS 5+ /Ubuntu
Database	PostgreSQL 8.4 and PostGIS 1.5
JDK	1.6 +
Proj	4.6.1
GEOS	3.2.0
libxml	2.7.6
Notes: PostgreSQL and PostGIS is the database for GEOSS	

Clearinghouse; Proj, GEOS and libxml are needed for PostgGIS installtion

Operational Requirements

Table 2. Operational requirement for GEOSS Clearinghouse

Specification	Minimum requirement
CPU core number	1
CPU speed	N/A
Memory	1 GB

Image type (RAM, local disk)

Table 3. Image information of GEOSS Clearinghouse

Image Id	Platform	Image size	Visibility	Description
ami-48bf4421	Cent OS, 64 bit	10 GB	Private	centos56_20110624_1

Table 3. Instance information of GEOSS Clearinghouse

Instance Id	Instance type	Memory	CPU Cores	Storage	platform
i-4131fc20	Large standard instance(m1.large)	7.5 GB	2	850 GB	64bit

Data storage

Currently there are about 80 K metadata in the GeoCloud. These metadata are stored in PostGIS database and the storage is about 2.8 G. An EBS volume with 50G is mounted on the system as the data storage.

Upload monthly

N/A

Download monthly

N/A

Redundancy and Load balancing

Two EBS volumes with 50G are created and attached to the running node. One EBS volume is to keep the code and the other is mounted to the data directory. Such a separate volume has two benefits: a) one is to restore the GEOSS Clearinghouse system from the volumes in case the current head node instance crashes, and b) another is that the volume could be any size from 1 GB to 1 TB in size. As the GEOSS Clearinghouse is a data intensive application, hundreds or thousands of data records could be harvested. Therefore, such an EBS data volume would be perfect to resolve the storage capability problems.

In addition, the load balance and scalability capability have been configured and tested through the CloudFormation of AWS.

Deployment in the Cloud

Using existing images

The Figure 1 shows how to deploy GEOSS Clearinghouse onto Amazon EC2 platform.

Appendix A.1 shows the Installation and Configuration of GEOSS Clearinghouse.

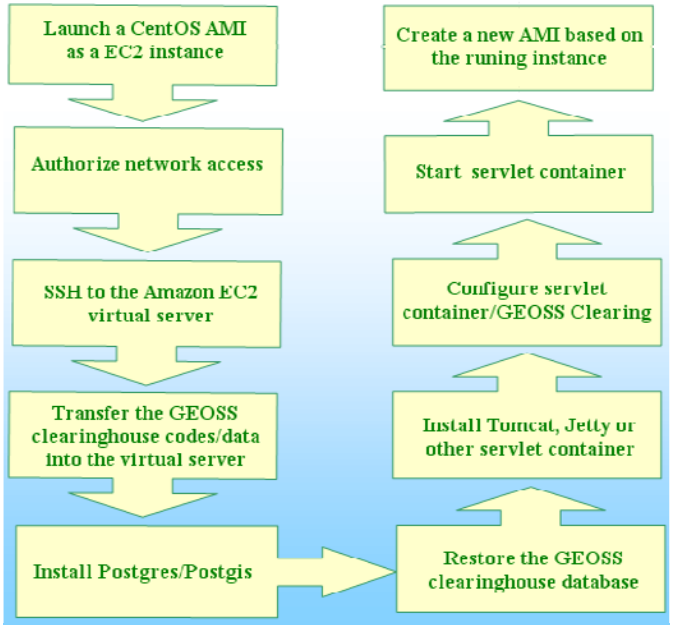

Figure 1 The deployment of GEOSS Clearinghouse onto Amazon EC2

Loading application with data

Appendix A.2 shows how to Migrate GEOSS Data to an Attachable EBS Volume This one-time process creates and redirects PostGRES to a persistent EBS volume. It's currently tested to work after the "Installing the GEOSS Clearinghouse dependent packages" step.

Customizing application suite

This section documents how to put scripts in place to automatically attach and mount a detachable EBS data volume at AMI instance startup, and scripts to start and stop the GeoNetwork application. Appendix A.1.3 shows the details of how to customize application GEOSS Clearinghouse with the scripts.

Operations in the Cloud

Monitoring of operations

AWS provides several matrix monitoring information, e.g., CPU unitization, network, memory, disk information. Through the AWS console, all those information can be retrieved.

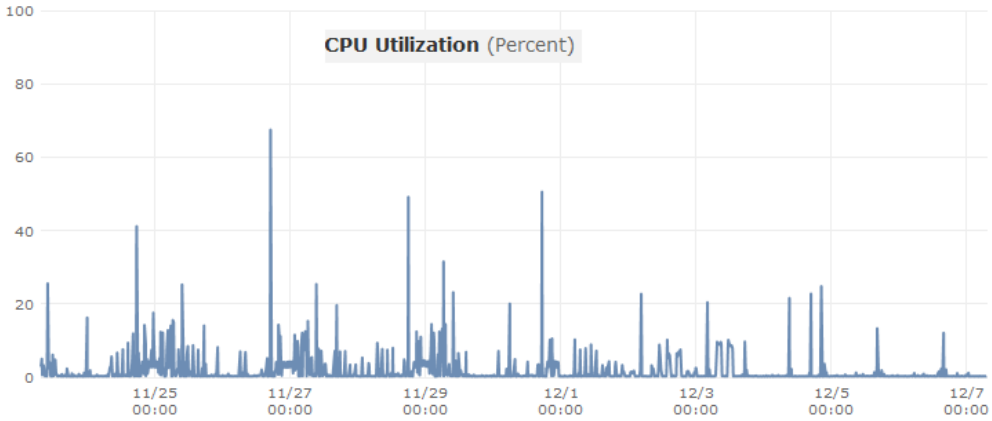

Figure 2. CPU average unitization with 5 minute as interval from 2011.11.5 to 2011.12.07

Figur 2 shows the CPU average unitization with 5 minute as interval from 2011.11.5 to 2011.12.07 minute. The results indicate that the CPU utilizations are usually below 20%.

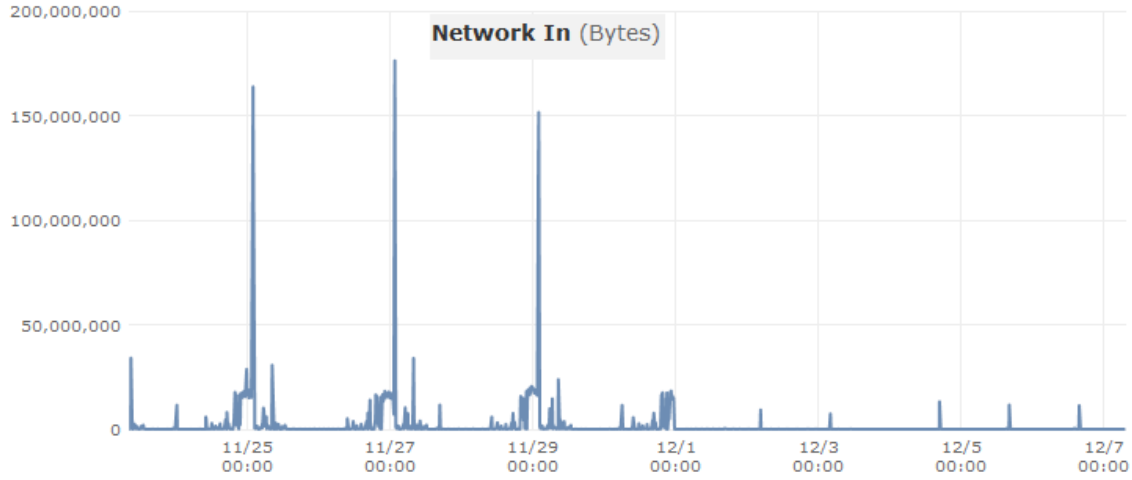

Figure 3. Average data in bytes transferred from the network with 5 minute as interval from 2011.11.5 to 2011.12.07

Figur 3 shows the average data in bytes transferred from the network with 5 minute as interval from 2011.11.5 to 2011.12.07 minute. The results indicate that the average data transferring from the network to the system is below 20 MB, and the system would have high data coming in when it is performing harvesting.

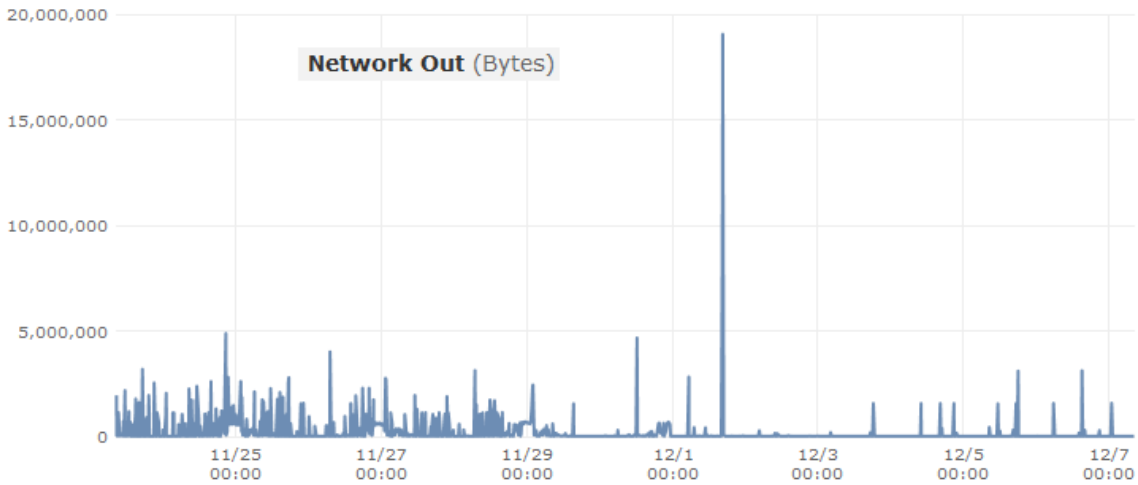

Figure 4. Average data in bytes transferred to the network with 5 minute as interval from 2011.11.5 to 2011.12.07

Figure 4 shows the average data in bytes transferred to the network with 5 minute as interval from 2011.11.5 to 2011.12.07 minute. T The results indicate that the average data transferring to the network to the system is below 50 MB. The results also indicate the stable access amount for the system.

Monthly usage, costs(tables and charts)

From the log files in GEOSS Clearinghouse cloud, we find that the usage in July, August and September are fewer than 10 K. Because the official GEOSS Clearinghouse moved to cloud and the European GEOSS portal connected to the cloud at the mid of the October, the usage increases to 17 K in October and 39 K in November.

Table 5. Monthly usage of GEOSS Clearinghouse

Month	Total(k)
July	2024
August	3896
September	17010
October	39090

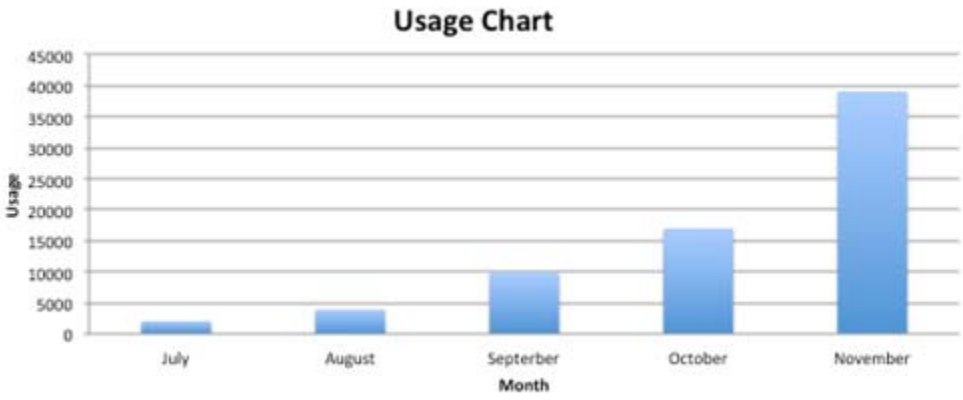

Figure 5. Usage chart from July to November, 2011

We launched the GEOSS Clearinghouse cloud at July 18. The Cost of July is 113.73 dollars, and the cost in August, September and October is about 270 dollars. Hence, the average cost of one month is about 270 dollars.

Table 6. Monthly Costs of AWS services

Month	Total(Dollar)	Amazon EC2		Amazon EBS	AWS Data Transfer
		Hours	Costs		
July	113.73	320	108.80	4.64	0.01
August	278.74	758	257.72	20.99	0.03
September	267.25	720	244.80	22.4	0.06
October	276.82	744	252.96	22.21	1.64

Discussion

Based on the monitoring information about the system performance, usage and costs, it is observed that : 1) the CPU utilization is around 20% , peaking to 80% at some point, which may be harvesting the records and handling the user access requests;2) the system usage are becoming larger with the time going by; 3) the cost of computing, storage and data transferring are very stable for each month.

Operational cost comparison (extrapolate to one year)

Fixed (one-time) costs versus monthly costs

Table 7. Cost comparison between AWS services and local server

	Cost		Amazon EC2	Local Server
Fixed (one-time) cost			None	Around $2000 to purchase the computing instances
Monthly	Network cost		$1 for data transferring / month	None (Included in the maintenance fee)
	Storage		$22 /month	None (Included in the first purchase)
	Computing power	Not reserved	$255 / month	~$17 (Assuming the server can be used for 10 years, and then the each month cost would be (2000/(10*12)))
		Reserved	$76 / month	
	Maintenance cost (Cooling, system, sys admin maintenance, room etc)		None	$200 / per month (Assuming $100 for cooling, network, and room fee, and $100 for paying a sys admin to check and maintain the server)
Yearly	Total	Not reserved	$3300	$2604
		Reserved	$1104	
Note: The cloud computing cost is based on the standard large instance with 2 CPU cores of EC2 to host GEOSS Clearinghouse. If more advanced instance types are selected, the computing power cost would increase				

Telecommunications

Twice telecom per month.

Operations and maintenance support

The cloud based GEOSS has been officially moved to the Amazon EC2. The web page link of old local server is redirected to the cloud server. Within the cloud, no extra maintenance for the cloud server is necessary. The daily maintenance and update for the website is required.

Security plan development and approval

AWS have taken the different approaches to secure the AWS infrastructure (http://aws.amazon.com/security/), including certifications and accreditations, physical Security, security services, protecting data privacy by encrypting data within the AWS cloud.

Different project has a specific product credentials. The product credentials will enable to carry out most of Amazon EC2 and S3 operations, including starting and stopping/terminating instances, creating, deleting and attaching volumes and snapshots, and creating and deleting Elastic IP addresses. Most project work should center on starting, stopping and terminating instances.

For the security reason, each product credential is not able to operate EC2 and volumes creating by other project credentials, and has not permissions to build images, and set up firewalls.

Issues and Lessons Learned

Security approval process

Currently, each product credentials encrypted with a public key (such as, gpg) and transferred over the internet. It would be great to have a process or method to enable each user to create/update its own credentials.

Recommendations on C&A

To implement the load balance and scalability, the platform services, such as the CloudFormation, and CloudWatch, better to be authorized for each product credentials.

Software deployment

Current, most of the projects share the same basic image, with pullulated software and library installed, such as, mysql, postgres, gcc, java, and tomcat. However, this may increase the burden of each system in the aspects of both storage and performance with more software than required installed and running.

Time-to-deploy

For the first time deployment, it takes around 1) 1 to 2 weeks to get familiar with the AWS services, 2) 2 days to customize the basic system running from a public machine image or hardened image, 3) 2 days to scripting the process of customizing the applications, 4) 1 week to explore the load balancer, auto-scaling capabilities, 5) others, elastic IP, data and code backup and recovery.

To get familiar with the AWS takes a long time for a non export, it might be better to have an introduction workshop for those new to the AWS to get a quick startup.

Failover, redundancy

When the concurrent number is larger than 600, there will be some failed response. Multiple copies can be distributed a different cloud regions for further protection of the system crash. The data can be backed up at a certain time interval.

Project planned future environment

In future, the system will be continuing to host under the cloud environment.

Appendix A. Documentation

A.1 GEOSS Installation and Configuration

Detailed installation and configuration instructions are available from the GeoCloud community portal.