GeoCloud Project Report – U.S. Census Bureau 2009 TIGER/Line Data Application

Description of Application

Operating Organization

The U.S. Census Bureau Geography Division (GEO) participated in the FY 2011 Federal Geographic Data Committee (FGDC) GeoCloud Sandbox Initiative by having the 2009 Topologically Integrated Geographic Encoding and Referencing (TIGER)/Line® data and download application hosted in the cloud.

TIGER/Line®Shapefiles are spatial extracts from the Census Bureau's TIGER database, containing features such as roads, railroads, rivers, as well as legal and statistical geographic areas. Currently the Census Bureau offers the datasets to the public for download from a Census data server.

Our objectives for the project were to evaluate the performance, cost, and security aspects of hosting TIGER/Line data in a cloud computing environment.

Community of Interest

The community of interest is the Census Bureau, the Geography Division, and the public TIGER/Line Shapefile user community.

The public TIGER/Line Shapefile user community consists of private companies; students; researchers; tribal, state, and local governments, and the public at large.

Operating System and Software Requirements

A Linux operating system (OS) and support for the following software was required:

Perl
Apache Web Server
SQLite
Zip
Carp
CGI
DBI
Template
YAML

Operational Requirements

The Census Bureau normally hosts the 2009 TIGER/Line data and download application in-house with a single virtual server housed at Bowie Computer Center and is one of numerous systems managed by the Computer Service Division (CSvD) staff. The in-house host server utilizes a RedHat Linux OS. The virtual host has access to 493 GB of storage (one of the mounts) that is shared between multiple GEO-applications and products besides the 2009 TIGER/Line data and download application.

The 2009 TIGER/Line data takes 56 GB of storage space. This data is rendered to end users via one instance of Apache httpd demon and vsftpd instance forFile Transfer Protocol (FTP), which runs on the same host. Optionally, the data can be accessed via custom applications, which interface with Apache web server through Common Gateway Interface (CGI) protocol.

The Census host server has the following hardware configuration**:**

- **CPU:** there are four (4) processors with model name "Intel(R) Xeon(TM) CPU 3.60GHz"
- **Memory:** 8GB
- **Disc Space**:2009 TIGER/Line data and download application resides on file system that has 493GB volume
- **Operating System**:RedHat Linux operating system 4.8.

RedHat Linux has versions (including current version): Update 8, 2009-05-18 (kernel 2.6.9-89The server has as much as 1000Mb/s of available bandwidth to the Internet, though not in a single transfer stream, and that capacity is used by all internal and external network use.

All traditional methods of downloading data are supported. The different modes of downloading the data are represented by the following methods and URLs.

- via anonymous ftp: ftp://ftp2.census.gov/geo/tiger/TIGER2009/
- via http (direct access): http://www2.census.gov/geo/tiger/TIGER2009/
- via http (CGI application): http://www2.census.gov/cgi-bin/shapefiles2009/national-files

The following chart lists the volume of transferred data from the AWS cloud instance of the 2009 TIGER/Line data and download application.

	June	July	August	September
Upload	-	-	-	-
Download	-	289.70 GB	214.29 GB	260.29 GB
Elastic IP	-	-	-	-

The following is the AWS cost associated with volume of upload, download data as well as usage of Elastic IP.

	June	July	August	September
Upload	0.0	0.0	0.0	0.0
Download	0.05	37.67	0.05	33.63
Elastic IP	0.17	0.0	0.17	0.0

Deployment in the Cloud

The Census Bureau management and Census GeoCloud team determined that is was feasible to participate in GeoCloud Sandbox Initiative due to low risk considering security and business operations as well as significant overall benefits. Benefits included gaining experience with the cloud environment (OS development and hardening, system configuration, application deployment, and administration), and obtaining actual data on the cost, performance, and reliability of cloud services.

The Census Bureau created a hardened image of the CentOS 5.x OS that was used by the Census Bureau and made available to the GeoCloud Sandbox Initiative participants who needed a Linux-based OS to support their applications. The CentOS was hardened to the CIS RHEL 5 Benchmark in order to comply with the Census Bureau Information Technology Security Policies and Procedures.

The OS was modified to make configuration persistent. These custom configurations included automatic startup of basic accounts, vsftpd, and httpd demons. These demons needed to read application data from persistent storage (EBS Volume) than from default location.

For our sandbox testing iterations of the OS build and patch and application installations, we used the ultra-cheap micro instance.

The final AWS configuration for the 2009 TIGER/Line data and download application was as follows

Large Instance

- 7.5 GB memory
- 4 EC2 Compute Units (2 virtual cores with 2 EC2 Compute Units each)
- 80 GB instance storage
- 64-bit platform
- I/O Performance: High
- API name: m1.large

Required Resources (Resources needed for TIGER/Line data and download application and its supporting software)		
	AWS	**Census Bureau**
Storage space	56 GB	56 GB
Storage Type		
Run time memory	7.5 GB	8GB
Operating System Type/Version	CentOS 5.6, 2.6.21.7-2.fc8xen #1 SMP	2.6.9-89.29.1.ELsmp #1 SMP Fri Sep 24 05:16:39 EDT 2010 x86_64 x86_64x86_64 GNU/Linux
CPU Type	Intel(R) Xeon(R) CPU E5430 @ 2.66GHz	Intel(R) Xeon(TM) CPU 3.60GHz
Number of CPU	2	4
Server Type With AWS, rate varies for micro vs. small vs. large instances	Large	

The AWS environment could be configured for auto scaling and load balancing. Auto-scaling could provide the ability to add system-resources automatically as needed, based on threshold configuration for memory, CPU, and bandwidth. The Census Bureau did not use the capability because it would not have worked for FTP protocol. In addition, the demand for system-resource was not anticipated to fluctuate significantly over time.

The EC2 instances were setup in multiple geographic zones for redundancy and security.

Using existing images

The Census Bureau used the hardened image of CentOS 5.x OS created by the Census Bureau for its own use and for use by GeoCloud Sandbox Initiative participants.

Loading application with data

The application and data were loaded simultaneously by creating a ZIP file on the original Census web server (www2.census.gov) which was then dropped into place on the Amazon cloud.

Customizing application suite

The TIGER/Line code needed almost no changes to work in the cloud. There were a few places where URLs needed to be changed to tigerline.census.gov, but for the most part no customization was needed. Apache needed to be configured to run tigerline.census.gov as it's primary hostname (rather than the various AWS IP addresses). Since the TIGER/Line data and application were in the tigerline.census.gov document root, there was not much Apache configuration required beyond this.

Installation scripting

The project produced reusable scripts and configuration as follows:

- Boot-strapping initiated necessary services and applications automatically when server started
- OS hardening: for customization of system security
- Configuration of initial user accounts enabled user accounts to automatically activate at system start-up

The deployment script was executed in CloudFormation.

Operations in the Cloud

Monitoring of Operations

The following aspects of the cloud instance were monitored and documented:

- Public Use of the Website - GEO staff visited the website daily and documented functionality in a weekly/monthly spreadsheet
- Data Integrity - script on AWS server checked for any changes to data (file size, date stamp), which was automated to send report to a Lotus Notes mail in database
- System Monitoring - CSvD conducted hourly monitoring of http, FTP,SSH and site availability monitoring, and measuring responsiveness. Monitoring notifications were emailed via Lotus Notes and designated staff were notified of significant problems via telephone call
- Additional monitoring was done on the AWS instance with monit to watch local processes, memory, CPU, disk, and track files. When there was high load on the system, it generated mail messages. If the checksum for integrity

failed, a message was sent using monit. We were unable to use HP OpenView tools because the connections through the firewall were not permitted.

Monthly Usage, Costs,(Tables and Charts)

Amazon Elastic Compute Cloud US East (Northern Virginia) Region		June	July	August	September
Amazon EC2 running Linux/UNIX	$0.34 per Large Instance (m1.large)	75.14	207.4	252.96	244
	$0.02 per Micro Instance (t1.micro)	1.22			
Elastic IP Addresses		0.17			
Amazon EC2 EBS		19.64	22.71	29.42	29.27
Amazon Simple Storage Service	US Standard Region	0	0	0	0
AWS Data Transfer (excluding Amazon CloudFront)		0.05	37.67	45.95	33.63
	Total	**96.22**	**267.78**	**328.33**	**306.9**

The following chart lists the volume of transferred data from the AWS cloud instance of the 2009 TIGER/Line data and download application.

	June	July	August	September
Upload	-	-	-	-
Download	-	289.70 GB	214.29 GB	260.29 GB
Elastic IP	-	-	-	-

The following is the AWS cost associated with volume of upload, download data as well as usage of Elastic IP.

	June	July	August	September
Upload	0.0	0.0	0.0	0.0
Download	0.05	37.67	0.05	33.63
Elastic IP	0.17	0.0	0.17	0.0

Operational Cost Comparison (Extrapolate to One Year)

For the purpose of comparing the financial costs, actual invoices are used for 2009 TIGER/Line on AWS during the months of June, July, August, and September of 2011. On the other hand, the cost of deploying TIGER/Line on Census Bureau platform is calculated assuming that a new server would be configured to host the deployment, in order to account for full hardware, OS, and support according to the Bureau's standard (see Table 2 below).

The capability of the AWS platform to provide extra resources dynamically was not utilized by the 2009 TIGER/Line data and download application. This was because the Internet traffic pattern that the product created during the trial period did not fluctuate significantly enough to trigger deployment of multiple instances.

Costs for cloud IaaS are substantially less when compared to the cost of supporting a dedicated server for the 2009 TIGER/Line data and download application. AWS cost averaged about $300 per month or $3600 per year. The estimated cost to maintain an in-house server for the application was $9483 for the first year and $7365 per year for the second year and beyond (year 2 and beyond do not include the one-time hardware cost).

In addition, offloading high-volume data transfers reduces the burden on our internal Internet connection, potentially saving hundreds of thousands of dollars in telecommunications costs by avoiding an expansion of additional bandwidth to support it. Our FTP2/WWW2 load runs about 90Mb/s at the 95th percentile and accounts for maybe 9TB per month of downloads. That's about $1,080/month, or $12,960/year. An extra 100Mb/s of bandwidth is in the $30K/MONTH range (using MTIPS), or $360K/year, which is very cost effective. This alone could make the case for hosting all our public data out on the cloud. The following table shows the itemized cost of standing up a new server to host 2009 TIGER/Line data and download application according to the Census Bureau standard.

Item	Cost
Hardware:	**Hardware Cost: $2118**
• 1x4 VM	$930
• .09TB	$1054
• tape	$135
Software:	**Software Cost: $3013**
• RHEL/ 5 year @$12495/5 years	$ $2499
• HPOV	$514
Services:	**Services Cost: $4352**
• support, medium	$4199
• storage	$153
Total Cost:	**$9483**

Table 2: Configuration of Host at the Census Bureau

Operations and Maintenance Support

The 2009 TIGER/Line data and download application did not require maintenance once it was deployed on AWS platform. The AWS platform provided the ability to expand available system resources if necessary, although it was not utilized.

Security Authorization to Use

The Census Bureau's Office of Information Security (OIS) attempted to leverage an existing Authorization To Operate by the Federal Risk Authorization Management Program (FedRAMP) or by General Services Administration (GSA) through the Apps.gov program. Doing so would have provided a strong foundation by which to determine a trust-based assessment of risk of the Infrastructure as a Service (IaaS). However, because FedRAMP was still in the planning phase, there was no such assessment conducted by FedRAMP or GSA on the Amazon Web Services.

While there was not a specific requirement for the IaaS to be pre-authorized in order for Census to participate, it was the expectation based on information provided by the FGDC GeoCloud team at the start of the project. The Census project team recognized that OIS would need to base their assessment and recommendation for authorization on the results of a C&A conducted by Treasury for the IaaS. Furthermore, OIS needed to ensure that the application layer including the operating system was built and configured securely to CIS standards in accordance with Census accepted baseline deviations.

Once it became apparent that FedRAMP and Apps.gov certifications would not occur during the GeoCloud Sandbox Initiative timeframe, the Census Bureau attempted to obtain a Service Level Agreement or enter into a Memorandum of Understanding (MOU) with Amazon AWS and eGlobalTech that would meet Federal Information Security Management Act (FISMA) requirements. Those efforts failed when Amazon could not accept our definition of a security breach, and could not accept the 5 day breach notification requirement. In addition, eGlobalTech could not accept the risk on behalf of Amazon AWS.

The Census Bureau Office of Information Security (OIS) ultimately made a recommendation to authorize the use of AWS. They relied on their risk assessment that took into account a SRA conducted by the Treasury Department on AWS and quarterly risk assessments provided by AWS.

OIS also determined that the risk of hosting the 2009 TIGER/Line data and download application in the cloud was considered low primarily due to the fact that the data and application were already publically available. The greatest concern was defacement of the website and insertion of illegitimate data.

The cloud instance of the 2009 TIGER/Line data and download application was released to the public when the Census Bureau Chief Information Officer (CIO) granted an Authority to Use (ATU) based on OIS's recommendation. The CIO agreed to accept the risk associated with the application hosted in the cloud provided we implemented a monitoring plan. The monitoring plan served to mitigate the risk and negated the need for a MOU. The plan consisted of monitoring and documenting aspects of the cloud instance. In the event of a security breach, access to the application would have been immediately directed back to the Census server.

Issues and Lessons Learned

The GeoCloud Sandbox Initiative validated some of the cost-reduction benefits of using external IaaS cloud computing services and demonstrated that the cost of supporting infrastructure can be eliminated by using external IaaS cloud computing services.

According to the observations and measurements made throughout the project, use of cloud technology offered by a public entity demands a certain level of effort to adopt. It presents a learning curve requiring knowledge transfer, new configuration, and development of scripts and software artifacts to automate computing processes, as well as organizational alignment to coordinate efforts both within the Census Bureau and with external partners.

In the meantime, cloud technology has and still is evolving rapidly. Management of the deployment environment has improved. It has become possible to version AWS configuration, new features have made deployment of content more efficient, and automated implementation of security rules has become possible (more). Additionally, cloud computing has become more ubiquitous, being offered for personal use (through services like iCloud).

These developments indicate that future use of the technology will be less demanding. On the other hand, the same rapid development of the technology that makes it more user friendly, also limits the time span during which the knowledge and reusable artifacts achieved and developed in the GeoCloud pilot project remain viable.

Additional observations regarding the overall use of cloud services are as follows:

Reliability:

- Cloud IaaS is a viable option for the TIGER/Line data and download application, and possibly other Census Bureau applications as the service did not experience disruptions and the team did not encounter technical difficulties with the pilot project.

Performance:

- The 2009 TIGER/Line data and download application did not see significant use during the test period. We need to test a more active application before reaching a definitive conclusion as to the performance of cloud IaaS.

Cost:

- Cloud IaaS offers significant cost savings related to hardware and telecommunications.

Scalability:

- Further assessment needs to be done to fully analyze the major benefit of cloud services which is scalability (rapid elasticity, easily expanding and contracting resources based on demand).

Ease of Configuration:

- The OS was successfully configured to meet our requirements (based on CIS benchmarks)
- One of the packaged AWS offerings (the "large instance") met the 2009 TIGER/Line data and application needs

Ease of Monitoring the System:

- AWS tools were sufficient to manage and analyze the basic status of the instance, but additional monitoring process were implemented in order to meet all our needs
- Web log files were accessible.

Security Approval Process and Recommendations on C&A

The greatest challenge for Census was attempting to get Amazon to meet Census Bureau security requirements in accordance with the FISMA. We had not anticipated the level of effort it took to attain a reasonable level of information assurance. However, we were able to attain that level of assurance due to the conclusions reached in the Risk Assessment that was conducted by the Office of Information Security (OIS).

The risk assessment focused on three primary security elements; Confidentiality, Integrity and Availability. The potential impact of this data being seen by anyone is extremely low due to the nature of the information being inherently mass disseminable, resulting in a very low confidentiality impact level. Since there was a newer data set that was published, the integrity of the 2009 data set wasn't crucial either. Furthermore, it was determined that reconstitution of the system internally was extremely fast and efficient resulting in effective compensating controls to any attack on the availability of the system within AWS. That analysis coupled with the low probability of attack on this type of mass disseminble information resulted in the conclusion that the overall risk associated with operating this system within AWS was very low.

Future participation by the Census Bureau in the FGDC GeoCloud Sandbox Initiative will be predicated on FedRAMP's authorization of a cloud environment hosted by Amazon. Ideally, Census would like to see a community cloud dedicated to the Federal government that complies with FISMA mandates. While we recognize there has been progress on this, Amazon will have to furnish an Authorization letter from the Joint Authorization Board (JAB) in order to comply with this requirement.

Time-to-Deploy

Due to reusable artifacts implemented in the pilot project, the estimated time needed for future deployments is reduced significantly. The time-to-deploy was 183 and 56 hours during the transition from development (sand-box) to production within AWS environment. Additionally, the time is projected to take 16 and 12 hours respectively.

Failover, Redundancy

AWS provided the capability to expand system resources as needed. According to the configuration, additional CPU and system memory could be made available if required.

Project Planned Future Environment

The 2009 TIGER/Line data and download application has been restored on our initial internal hosting environment.

The Census Bureau is in the process of implementingan internal cloud configuration to host data and applications in the near future.