Health Resources and Services Administration

Geospatial Data Warehouse:

Integrating Spatial and Tabular Extract, Transform, and Load Processes

ESRI 2005 International Users Conference
July 24-29, 2005

June 3, 2005

Health Resources and Services Administration
5600 Fishers Lane
Rockville, MD 20857

Authors:

Terri L. Cohen – Health Resources and Services Administration (HRSA)
Julie R. Baitty - Health Resources and Services Administration (HRSA)

Table of Contents

List of Figures

Health Resources and Services Administration Geospatial Data Warehouse: Integrating Spatial and Tabular Extract, Transform, and Load Processes

ABSTRACT

The Health Resources and Services Administration Geospatial Data Warehouse (HGDW) is designed to support the mission of the Health Resources and Services Administration (HRSA) by making tabular and spatial information available on HRSA supported programs and related health resources. This information is refreshed on a scheduled basis using Extract, Transform, and Load (ETL) processes developed using Informatica.

The HGDW extends traditional data warehouse ETL activities and processes by encompassing address correction and cleansing (via Trillium Software), and integrating ESRI's ArcGIS geoprocessing model technology to geocode and load spatially-enabled data into the HGDW databases. The entire spatial and tabular refresh process is driven through a single Informatica ETL workflow for each HRSA-specific data layer.

This paper describes the refresh process and discusses how the integration of these technologies has reduced the data warehouse data refresh time from weeks to just a few days.

INTRODUCTION

About HRSA

Vision

The Health Resources and Services Administration (HRSA) envisions optimal health for all, supported by a health care system that assures access to comprehensive, culturally competent, quality care.

Mission

HRSA provides national leadership, program resources and services needed to improve access to culturally competent, quality health care.

Goals

As the Nation's Access Agency, HRSA focuses on uninsured, underserved, and special needs populations in its goals and program activities.

About the HRSA Geospatial Data Warehouse

The HRSA Geospatial Data Warehouse (HGDW) and its associated applications were developed to provide a single point of access to a broad range of HRSA programmatic information, related health resources, and demographic data, to both HRSA and the general public. The information within the HGDW can be displayed as tables, reports and dynamically-generated maps. The overarching goal for the HGDW is to serve as a single source of information for reporting on HRSA's activities, and to promote information sharing and collaboration within and between HRSA, its partner agencies, State and local health planners and policy makers, and stakeholders.

The HGDW is accessible from the HRSA Web site at (http://www.hrsa.gov) or directly at http://datawarehouse.hrsa.gov. It can also be found in the *Human Health and Disease* channel on the Geospatial One-Stop portal (http://geodata.gov). This portal is part of the Geospatial One-Stop E-Government initiative that provides access to geospatial data and information. Additionally, the HGDW spatial metadata collection is registered with the Federal Geographic Data Committee (FGDC) Clearinghouse (http://fgdc.er.usgs.gov).

A key objective of the HGDW is to create an overall design that integrates traditional tabular data, in the form of data marts, with spatially-enabled data such as geocoded addresses, thematic layers, and the results of spatial analyses. Each type of data influences and enriches the other.

Focus of This Paper

This paper describes the development and evolution of the ETL processes used in the HGDW, from a loosely connected mixture of automated and manual processes and procedures to a set of integrated, completely automated processes that comprise the data refresh process. It discusses some of the challenges related to the development of the integrated ETL process, and highlights some of the benefits.

OBJECTIVES OF ETL PROCESS INTEGRATION

The objectives of integrating the tabular and spatial ETL processes in the HGDW were to:

- improve the timeliness of the data in the HGDW;
- replace time-consuming, labor-intensive, complex manual processes with automated equivalents;
- eliminate the chances that data errors would be introduced as a result of using manual processes; and
- make resources available to carry out other functions such as new development.

CHALLENGES

During the evolution and development of the integrated ETL processes, the HGDW encountered and solved several significant challenges. Among these were:

- to enhance the quality of address input data in order to improve the geocoding match rate and positional accuracy;
- learning to use ArcGIS Model Builder effectively and efficiently, including learning how to select the appropriate tools and how to structure processing flow through a complex model;
- ascertaining the correct method for implementing Python scripts including calls to ESRI Model Builder objects for execution from within Informatica as external procedures; and
- managing topology and attribute related issues when automating the ArcSDE load processes.

TECHNICAL ARCHITECTURE

Overview

The HGDW is comprised of several MS SQL Server 2000 databases that house a data staging area, multiple data marts, spatial data, and spatial and tabular metadata.

The geographic scope of the HGDW includes the United States and its Territories, in addition to some ancillary information on Mexico in support of HRSA's U.S. – Mexico Border Health Initiative. Within each spatial database, the data are broken into sets of feature classes by type of geography (State, county, census tract, block group, and so forth) and by locale (contiguous United States, Alaska, Hawaii, Puerto Rico, U.S. Virgin Islands, American Samoa, Federated States of Micronesia, Guam, Northern Mariana Islands, and the Republic of Palau).

Software Environment

The software environment that supports the HGDW is a combination of technologies. All of the servers operate using Microsoft Advanced Server 2000. The databases are all Microsoft SQL Server 2000. Spatial data are stored and managed using ArcSDE 9.0.1; spatial processing, model execution, and spatial data loading are performed using ArcInfo 9.0.1 and Python 2.1.

Tabular data are loaded using the Informatica ETL tool in the SQL Server 2000 data staging area on the Informatica server. Once the staged data have been prepared they are transferred to the main database server using Informatica ETL mappings that are part of the Informatica workflow being executed.

Data are made available to the application through direct queries of the SQL Server databases, as well as through ArcIMS map services and the Hyperion Performance Suite online analytical processing (OLAP) software tool. These products are all used downstream from the ETL process, and consequently do not play any role in it.

Figure 1 provides a diagrammatic representation of the HGDW architecture.

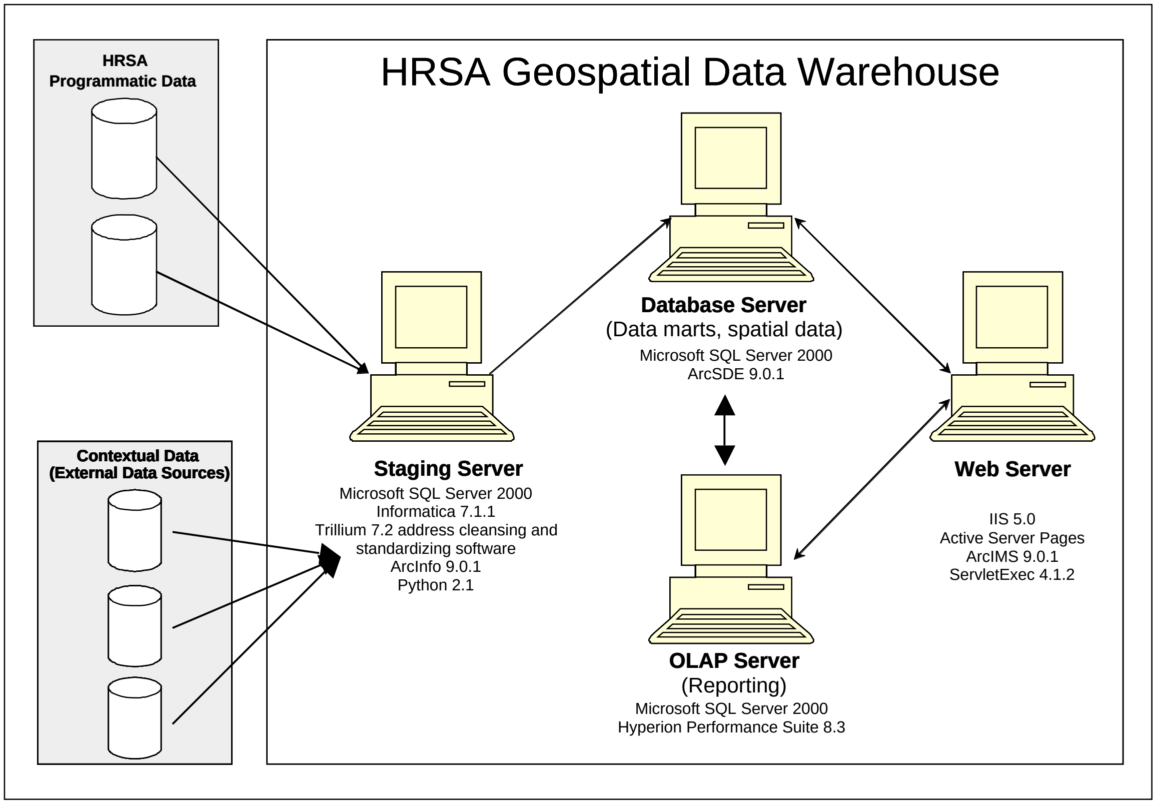

Figure 1 HRSA Geospatial Data Warehouse Technical Architecture

PROCESS FLOW DESCRIPTION

Data are extracted from the HRSA source systems, and are also obtained from sources external to HRSA on various electronic media. Figure 2 provides a schematic overview of the processing flow and decision tree. Each ETL process is controlled by an Informatica workflow. An Informatica workflow has the ability to perform automated transformations, in addition to being able to launch and respond to the results from external processes, such as the Trillium address cleansing software and the geoprocessing tasks implemented as Python scripts. It is through Informatica workflows that the processing steps are controlled; processing is suspended until successful completion of the external tasks.

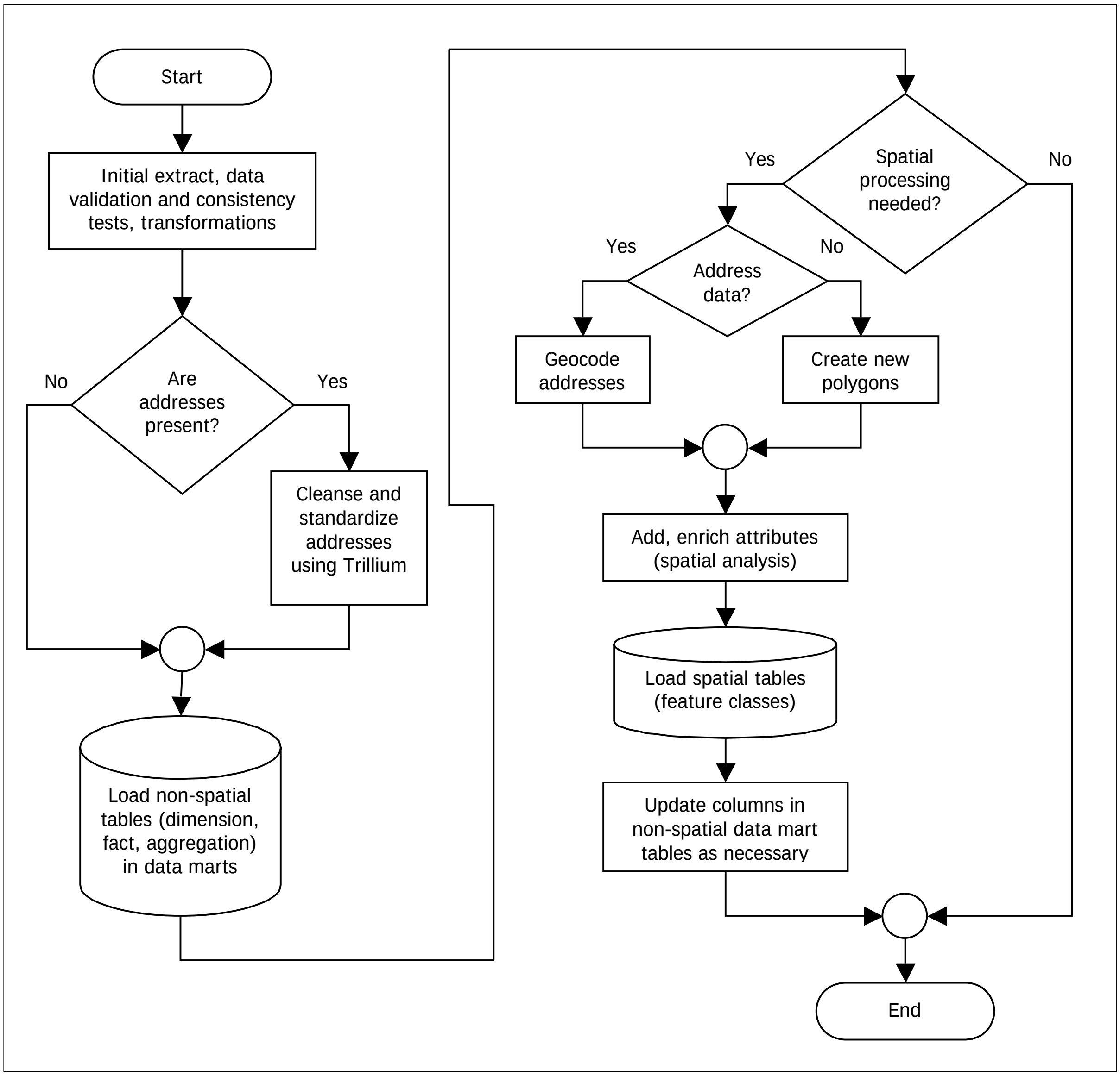
Start
Initial extract, data validation and consistency tests, transformations
Are addresses present?
No
Yes
Cleanse and standardize addresses using Trillium
Load non-spatial tables (dimension, fact, aggregation) in data marts
Spatial processing needed?
Yes
No
Address data?
Yes
No
Geocode addresses
Create new polygons
Add, enrich attributes (spatial analysis)
Load spatial tables (feature classes)
Update columns in non-spatial data mart tables as necessary
End

Figure 2 HGDW Spatial and Tabular ETL process flow

The ETL process begins by transferring the data into working tables in the data staging area. Data consistency and validity tests are performed at this time. Once the data are validated, data with an address component are run through the Trillium Software Data Quality Connector for Informatica PowerCenter. The final step in the ETL process for tabular data is to truncate and reload the fact tables, truncate or update related dimension tables, and create the aggregation tables. These tables all exist in the data marts, not in the staging area.

Upon completion of the tabular data loading, the ETL process follows one of two branches for tabular data that have a spatial component:

- Tables containing address data are geocoded using geoprocessing scripts to launch the ArcGIS StreetMap USA geocoder. After the address corrections are completed, the results are stored in a set of spatially-referenced tables in the data staging area.

- Datasets that require geoprocessing, other than geocoding, are handled by appropriate geoprocessing scripts. The results of the geoprocessing are distinct from the input data, and are managed as intermediate tables in the data staging area.

All HRSA-specific spatially-referenced data are processed to enrich their attribute sets, by performing point-in-polygon and/or polygon intersection spatial analyses.

Lastly, the spatial data are loaded into the feature classes used by the HGDW application, outside the data staging area. The feature classes and data marts reside on the same server, in separate SQL Server databases.

The HGDW performs a complete replacement of the existing spatial data at each refresh rather than selective additions, edits, or deletions based on reconciliation of the source and target data. At present the HGDW is not required to maintain row-level or cell-level change histories, which makes this approach feasible. It is simpler and less resource intensive to implement.

Data Cleansing and Validation

The Informatica workflow runs the data through a series of quality, validity, and consistency checks before initiating the next step. Invalid and inconsistent data are reported back to the source system data owners for correction before they are included in the HGDW. The workflow process outputs files that list failed records, in order to facilitate the data correction process.

Data containing address information are cleansed and standardized using the Trillium Software Data Quality Connector for Informatica PowerCenter, so that address correction according to the United States Postal Service postal address data base can occur. Address standardization helps to increase the consistency and accuracy of the attribute data.

Address standardization provides an additional feedback loop to the source systems. Addresses that are altered by the HGDW ETL process are reported back to the owners of the HRSA-source system databases so the originating data can be standardized and improved.

All the data cleansing and validation steps in the ETL process occur in the data staging area. The Informatica workflow being executed directly populates the relevant dimension, fact, and aggregate tables in the data marts.

Geoprocessing

After the data cleansing and validation is complete, geoprocessing is initiated as required. The ETL implements four types of geoprocessing:

- geocoding of address data;
- creating polygonal features by merging and dissolving features from other types of polygons;
- performing spatial analyses to enrich or derive additional attributes; and
- transferring the final results from the staging area to the final ArcSDE-managed feature classes.

The HGDW conducted extensive analysis and comparison of the geocoding and attribute enrichment that results from using the Trillium geocoding product versus those obtained by geocoding using StreetMap USA and subsequent spatial analysis. Based on the results of these comparisons, the HGDW elected to use the results from StreetMap USA and geoprocessing, despite the additional steps required. The reason for the differences in results arises from Trillium's origins as a tool primarily oriented to serving the needs of large-volume mailers, versus those related to digital mapping.

Trillium software determines county information based on the business office for the five digit ZIP code for an address; this is not always correct for addresses that are in ZIP codes that cross county lines. Where a nine-digit ZIP code is included with the address data, coordinates that represent the location of the address are obtained by interpolation from the data in the Trillium U.S. Rooftop Census Geocoder. However, all addresses that contain only a five-digit ZIP code are placed at the centroid of the ZIP code. This is in contrast to the interpolation method used by ESRI's Streetmap USA product, which only places points at ZIP code centroids when no suitable street segment is found in the reference data. Thus, there is a higher likelihood that significant location errors and attribute discrepancies may occur by using the Trillium-reported location and attributes for mapping applications as opposed to mailing.

In addition to obtaining attributes that will be more consistent with the results displayed on a map, other types of data enrichment are obtainable only through spatial analysis. These include derivation of the current county and Congressional District into which an address falls, and whether a point falls into, or a polygon intersects, the U.S. – Mexico Border Health Initiative area.

Each geoprocessing script is developed using ArcGIS Model Builder, and is then exported to a Python script. Each script is launched and run as external process by the controlling Informatica workflow.

Spatial Data Loading

Transferring data from the staging area to the final data mart tables is performed by the Informatica workflow prior to any geoprocessing. The geoprocessing scripts transfer the spatial data from the staging area to the production feature classes. The final update of tables in the data marts using the spatially-derived attributes is handled by the Informatica workflow, once the geoprocessing script has completed populating the feature classes.

A significant challenge encountered during development of the integrated ETL process was to reduce the probability that topological errors would prevent complete loading of the final spatial results into the feature classes in the database. Each geoprocessing script applies a set of topological rules and corrections to the staged data before the transfer to the final destination tables is attempted. These rules include verifying that polygonal data are free of gaps, overlaps, and sliver polygons; that they do not contain short segments, and that they are not self-intersecting.

Due to the design of the HGDW mapping application, it is important that the individual ArcSDE-assigned layer id for each layer be maintained across time. This constraint puts an extra layer of complexity on the geoprocessing step that performs the final loading. It is not operationally desirable to drop and recreate feature classes at each data refresh; the existing feature classes must be truncated and then reloaded using the Append tool from ArcToolbox. Thus, it is also essential that the set of attributes on each layer be stable. The Append tool will not properly load data when either the attribute sets or definitions are different between the source and target feature classes.

Benefits of Integrating Spatial and Tabular ETL Processes

Several direct benefits have been realized by creating an integrated ETL process for both spatial and non-spatial data in the HGDW. These include:

- facilitating more frequent refresh from source systems, by reducing the amount of time required to perform an update;
- reducing processing errors and inconsistencies through automating and standardizing manually processes;
- eliminating the data refresh induced time lag between spatial and non-spatial data; and
- enabling reallocation of resources to other areas of HGDW development.

NEXT STEPS

At present, the HGDW operates on a quarterly refresh schedule on most HRSA-specific spatial and tabular data sets. In the near future, this schedule is planned to increase in frequency to weekly or daily for key datasets. In addition, the HGDW may be required to maintain a history of the data. Altering the ETL processes to meet these changing requirements will pose the next series of challenges.

ACKNOWLEDGMENTS

The HGDW Team would like to acknowledge Dr. Elizabeth M. Duke, HRSA Administrator, for her continued support of this project.

AUTHORS

Terri L. Cohen
Health Resources and Services Administration (HRSA)
Office of Information Technology
Parklawn Bldg., 5600 Fishers Lane, Rm. 10A-30
Rockville, MD 20857
tcohen@hrsa.gov

Julie R. Baitty
Health Resources and Services Administration (HRSA)
Office of Information Technology
Parklawn Bldg., 5600 Fishers Lane, Rm. 10A-30
Rockville, MD 20857
jbaitty@hrsa.gov

Health Resources and Services Administration
5600 Fishers Lane
Rockville, MD 20857