



United States
Department of
Agriculture

Service Center
Implementation
(SCI)

Implementation of Geospatial Data Warehouses II

*Prepared by
Science Applications International Corporation (SAIC)
For the
Service Center Initiative, Data Management Team*

Abstract:

The Implementation of Geospatial Data Warehouses II provides the business case and mandate for the implementation of a USDA Geospatial Data Warehouse. The essential components of the Geospatial Data Warehouse are described and an implementation plan is established.

Keywords:

Data Warehousing, Geospatial, GIS, Data Mart, Metadata

Introduction

The Service Center Initiative (SCI) Data Management Team #5: Geospatial Data Standards developed the February 2002 *Implementation of Geospatial Data Warehouses II* to illustrate the USDA vision of how geospatial data warehousing will be used implemented and used for on-line data access and data dissemination in the near and long-term.

The individuals who contributed to the development of this document are:

- David Anderson, (NRCS) Service Center Data Team Leader
- NRCS National Cartography and Geospatial Center (NCGC) Staff - Tommie Parham, Director
- Farm Service Agency (FSA) Air Photo Field Office (APFO) Staff - Ron Nicholls, Director
- NRCS Information Technology Center (ITC) Staff – Jack Carlson, Director
- NRCS National Water & Climate Center (NWCC) - Rick Roberson, Team Leader
- Science Applications International Corporation (SAIC) – Tom McCarty, Nicole Soltyka

Table of Contents

TABLE OF CONTENTS	II
LIST OF FIGURES	IV
LIST OF TABLES	IV
LIST OF ACRONYMS	V
EXECUTIVE SUMMARY – GEOSPATIAL DATA WAREHOUSE IMPLEMENTATION.....	1
BACKGROUND	1
C) ACCESS - ACCESS TO THE DATA WAREHOUSE IS PROVIDED THROUGH SEAMLESS INTERFACES IS THE COMPONENT THAT IS SEEN BY MOST OF THE USERS. THIS COMPONENT CONTRIBUTES MUCH OF THE BUSINESS CASE BENEFITS OF GEOSPATIAL DATA WAREHOUSE. SEAMLESS, EASY-TO-USE INTERFACES WILL CONTRIBUTE TO A SUCCESSFUL WAREHOUSE STRATEGY. USERS WILL WANT TO ADD TO WAREHOUSES IF THE DATA CAN BE EASILY LOADED AND RETRIEVED FOR BUSINESS NEEDS.....	4
IMPLEMENTATION PLAN.....	4
SUMMARY	5
1. BACKGROUND	6
1.1. DISTRIBUTED MANAGEMENT OF GEOSPATIAL DATA.....	6
1.2. REQUIREMENTS-BASED ACCESS AND DISTRIBUTION MODELS.....	7
1.3. WAREHOUSE IMPLEMENTATION STRATEGY AND PROTOTYPING.....	7
2. SCOPE AND OBJECTIVE.....	8
3. GEOSPATIAL DATA WAREHOUSES BUSINESS CASE.....	8
3.1. SECRETARY OF AGRICULTURE FOOD AND AGRICULTURE POLICY REPORT	9
3.2. E-GOV INITIATIVE.....	9
3.3. SECURITY.....	10
4. GEOSPATIAL DATA WAREHOUSE REQUIREMENTS	11
4.1. BUSINESS REQUIREMENTS.....	12
4.2. SYSTEM REQUIREMENTS.....	14
4.3. GEOSPATIAL DATA WAREHOUSE VISION (TECHNICAL ARCHITECTURE).....	14
4.4. BUILDING A DATA WAREHOUSE FOR THE “VISION”	14
4.5. ARCHITECTURE ALTERNATIVES	16
4.6. ACQUISITION AND INTEGRATION.....	20
4.7. STORE AND MANAGE.....	23
4.8. ACCESS	27
5. CURRENT DATA CENTER ARCHITECTURE	28
5.1. APFO	29
5.2. NCGC – FORT WORTH, TX	29
5.3. NRCS ITC – FORT COLLINS, CO.....	29
5.4. NWCC – OREGON	29
5.5. FSA – KANSAS CITY, MO.....	29
6. USDA INFORMATION TECHNOLOGY ORGANIZATIONS.....	30

6.1.	COMMON COMPUTING ENVIRONMENT (CCE).....	30
6.2.	ELECTRONIC ACCESS INITIATIVE (EAI)	30
7.	EXTERNAL IMPACTS AND INFLUENCING TECHNOLOGIES.....	30
7.1.	NSDI	31
7.2.	OPENGIS	31
7.3.	GEOSPATIAL ONE-STOP	32
7.4.	ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE (ESRI).....	32
7.5.	OTHER GOVERNMENTAL AND COMMERCIAL GEOSPATIAL DATA WAREHOUSES.....	33
8.	TECHNOLOGY REVIEW	34
8.1.	AVAILABILITY	34
8.2.	WEB SERVICES.....	38
9.	IMPLEMENTATION PLAN – A PHASED APPROACH.....	39
9.1.	NEAR-TERM IMPLEMENTATION PHASE.....	39
9.2.	GEOSPATIAL DATA WAREHOUSE SUPPORT SERVICE ROLES AND RESPONSIBILITIES.....	42
9.3.	GEOSPATIAL WAREHOUSE PILOT SCHEDULE AND COST	44
9.4.	LONG-TERM IMPLEMENTATION PHASES.....	48
10.	SUMMARY.....	49
	APPENDIX A - CURRENT DATA CENTERS ARCHITECTURE.....	50
	APFO ARCHITECTURE.....	50
	HARDWARE.....	50
	SOFTWARE.....	50
	TELECOMMUNICATIONS.....	50
	NCGC – FORT WORTH, TX.....	51
	HARDWARE.....	51
	SOFTWARE.....	51
	TELECOMMUNICATIONS.....	51
	SECURITY.....	51
	DATA.....	51
	STAFF.....	52
	NWCC – OREGON	52
	HARDWARE.....	52
	SOFTWARE.....	53
	TELECOMMUNICATIONS.....	53
	SECURITY.....	53
	DATA.....	53
	STAFF.....	54
	CURRENT CCE ARCHITECTURES AND IMPACTS	54
	INTERMEDIATE DATA ARCHITECTURE.....	54
	LONG-TERM DATA ARCHITECTURE.....	54
	RECOMMENDED TECHNICAL ARCHITECTURE.....	55
	CURRENT EAI ARCHITECTURE AND IMPACTS	56
	OCIO STACK(S)	56
	EXTERNAL WEB FARM (WFE).....	57
	INTERNAL WEB FARM (WFI)	57
	TEST/CERTIFICATION WEB FARM (WFTC)	58

STAFFING.....	58
BANDWIDTH.....	58
APPENDIX B – REPLICATION OPTIONS - GARTNER GROUP REPORT	59
APPENDIX C – BIBLIOGRAPHY	60

List of Figures

Figure 4-1 Current Geospatial Data Management Environment	16
Figure 4-2 USDA Geospatial Data Warehouse Architecture Example	19
Figure 4-3 USDA Abstract Data Warehousing Architecture	20
Figure 4-4 PLSS Data Anomalies	22
Figure 4-5 SSURGO Tabular Data Structure Diagram	24
Figure 7-1 OpenGIS technology to support the USDA Data Warehouse	32
Figure 7-2 EnviroMapper Interface	33
Figure 9-1 Pilot Geospatial Data Warehouse Implementation Schedule	45
Appendix A: Figure 0-1 Sample Hybrid Application Architecture	55
Appendix A: Figure 0-1 Web Farm Components in Relation to the USDA Network	56

List of Tables

Table 8-1 Alternative Replication Technologies	37
Table 9-1 Business Driver to Technical Requirement to Pilot Crosswalk	39
Table 9-2 Geospatial Data Warehouse Roles and Responsibilities	42
Table 9-3 Pilot Geospatial Data Warehouse Support Service Labor Costs	45
Table 9-4 Phase I Geospatial Data Warehouse Support Service Labor Costs.....	46
Table 9-5 Phase II Geospatial Data Warehouse Support Service Labor Costs	46
Table 9-6 Geospatial Data Warehouse (Pilot - Phase II) Budget	48

List of Acronyms

APFO	Aerial Photography Field Office
AWDB	
CBA	Cost Benefit Analysis
CCE	Common Computing Environment
CDBS	Centralized Database System
CDF	Centralized Forecasting System
CDROM	Compact Disk – Read Only Memory
COM	Common Object Model
CRP	Conservation Reserve Program
CVS	Concurrent Versions System
DBMS	Data Base Management System
DNS	Domain Name System
DOQ	Digital Ortho Photo
DRG	Digital Raster Graph
DSS	Decision Support System
EAI	Electronic Access Initiative
EPA	Environmental Protection Agency
ESRI	Environmental Systems Research Institute
ETL	Extract, Transform, Load
FC	Fort Collins
FGDC	Federal Geographic Data Committee
FSA	Farm Service Agency
FTP	File Transfer Protocol
GIS	Geographic Information Systems
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer Protocol
IP	Internet Protocol
IT	Information Technology
ITC	Information Technology Center
ITWG	Information Technology Working Group
KB	Kilobyte
KC	Kansas City
LWV	LAN/WAN/Voice
Mbps	Mega bits per second
MDOQ	Mosaicked Digital Ortho Quad
MIT	Massachusetts Institute of Technology
MrSID	Multi-Resolution Seamless Image Database
NASIS	National Soils Information System
NCGC	National Cartography and Geospatial Center
NetCDF	
NRCS	Natural Resource Conservation Service
NRI	National Resources Inventory
NSDI	National Spatial Data Infrastructure
NWCC	National Weather and Climate Center

OCIO	Office of the Chief Information Officer
OGC	Open GIS Consortium
OLAP	On-line Analytical Processing
OLTP	On-line Transaction Processing
PC	Personal Computer
PKI	Public Key Infrastructure
PLSS	Public Land Survey System
PRISM	Parameter-elevation Regressions on Independent Slopes Model
RAID	Redundant Array of Independent Disks
RD	Rural Development
RDBMS	Relational Database Management System
RPO	Recovery Point Objectives
RTO	Recovery time objectives
RUSLE	Revised Universal Soil Loss Equation
SC	Service Center
SCA	Service Center Agencies
SCAN	Soil Climate Analysis Network
SCIMS	Service Center Information Management System
SCMI	Service Center Modernization Initiative
SCOTS	Standards-based Commercial Off-the-Shelf
SMTF	Simple Mail Transfer Protocol
SNOTEL	Soil Climate Analysis Network
SOAP	Simple Object Access Protocol
SRDF	Symmetrix Remote Data Facility
SSL	Secure Sockets Layer
SSURGO	Soil Survey Geographic Database
STL	Saint Louis
TB	Terra Byte
UCAN	Unified Climate Access Network
UDDI	Universal Description, Discovery, and Integration
USDA	United States Department of Agriculture
UUNET	
VLAN	Virtual Local Area Network
VPN	Virtual Private Network
WFe	External Web Farm
WFi	Internal Web Farm
WFtc	Test/Certification Web Farm
WSDL	Web Services Description Language
WYFOR	
WYSNO	
XML	eXtensible Markup Language

EXECUTIVE SUMMARY – GEOSPATIAL DATA WAREHOUSE IMPLEMENTATION

BACKGROUND

Concerns for Farm Bill Initiatives, Homeland Security, E-government, and the ongoing Service Center GIS implementation, all support the need to establish Geospatial Data Warehouses.

Acknowledging the importance of managing Geospatial data, in September 2000, the USDA Service Center Modernization Initiative, Data Management Team released a two-part document entitled *Phase I: Geospatial Data Management Requirements* and *Phase II: Implementation of Geospatial Data Warehouses*. The purpose of those documents was to evaluate and report on the business case and feasibility of USDA Geospatial Data Warehouse. The effort examined the business and technical requirements, architecture and staffing needs for the warehouse. The recommendations of this two-part study were as follows:

- Establish on-line data warehouses at Data Acquisition, Integration, and Delivery Centers
- House and support on-line web applications at USDA Web Farms.
- Establish a common Internet Portal as a “one-stop-shopping” service for access and delivery of geospatial data.

After this two-part study, USDA continued work on the business case and strategy for GIS overall in two documents entitled *Requirements And Cost Benefits Analysis: Managing Geospatial Data For Better Program Delivery* and *USDA Service Center Geographic Information Systems (GIS) Strategy*. The recommendations made in the cost-benefit analysis showed a favorable return-on-investment if USDA implemented a mixed centralized and distributed architecture for geospatial information management. One of the key pieces of this architecture framework is the existence of a Geospatial Data Warehouse component. To date these efforts have clearly defined the business case and vision for the management of geospatial information within the USDA Service Center enterprise.

PURPOSE

The purpose of this document is to layout the implementation plan for the Geospatial Data Warehouse in the context of the studies, acquisitions and implementation that have occurred over the past 3 years. The document provides specific details concerning the location of geospatial data, hardware and software, security, staffing and the telecommunications infrastructure that is required to implement and maintain an operational warehouse.

Currently geospatial data is loosely managed and consists mostly of a widely distributed collection of subject oriented spatial flat files, metadata and tabular data. In order to migrate from the status quo to an efficient, sustainable, cost effect, coherent and well-

defined data management environment, a common vision Geospatial Warehouses must be developed and implemented. The development and management of geospatial data in the manner described in *Implementation of Geospatial Data Warehouses II* is critical, whether the data is disseminated on media or via the web. Web mapping applications are becoming more important but are limited today by lack of current, highly reliable and available online geospatial warehouses. As GIS applications in the Service Center continue to develop, and as the bandwidth between Service Centers and Data Centers increases, web mapping will become more feasible.

BUSINESS CASE AND BENEFITS OF A GEOSPATIAL DATA WAREHOUSES

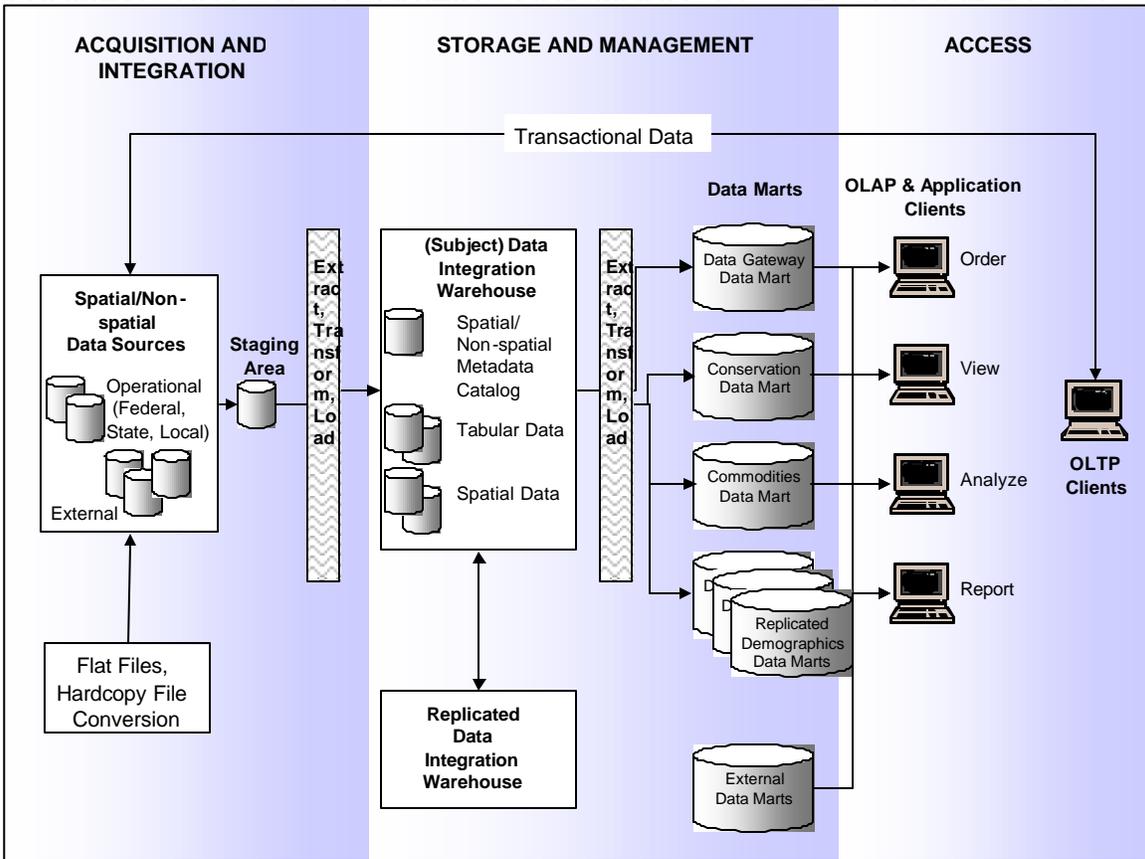
The business case for implementing the geospatial warehouse is strongly supported by the following major business requirements:

- Maintain an infrastructure that provides the authoritative source for integrated USDA geospatial data.
- Provide on-line access to national data sets to meet emerging program, Farm Bill and E-Gov requirements.
- Provide high availability of geospatial data.
- Provide a secure data access mechanism that also supports information privacy requirements.
- Provide a mechanism for USDA national, state and local data to be made available to the general public, conservation partners and third party vendors who support conservation planning.
- Enhance sharing of program and GIS data among Service Center Agencies.

The business case is supported by strong legislative mandates that state:

- Farm Bill – *Both the House and the Senate versions of the Farm Bill Provide for third parties to develop conservation practices using the Natural Resources Conservation Service technical information. This will create an increased level of urgency to get the Geospatial Warehouses operational.*
- USDA Food and Agriculture Policy, September 2001 – *GIS can help solve many policy and program problems by merging data sets from multiple agencies. Cross-agency data sharing and information integration technologies are needed to break down the stovepipes that thwart efforts to obtain efficient service.*
- E-Government, Geospatial One-Stop, OMB, Capital Asset Plan—*Geospatial One-Stop will revolutionize E-Government by providing a geographic component for use in all Internet based E-Government activities across local, state, tribal and Federal government.*
- FGDC Briefing on Homeland Security – *FGDC vision includes intent to encourage spatial data exchange at local, national, and global levels to help deal with social, economic, and environmental issues.*

FRAMEWORK FOR ORGANIZING AND MANAGING GEOSPATIAL DATA



a) Acquisition and Integration - The first major component of the framework is Acquisition and Integration. Much of this function already exists or is evolving within the two data centers, APFO and NCGC. This activity includes obtaining the data required for agency business processes and reformatting or reprocessing the data as needed according to approved standards. To support data warehousing requirements, the managers of these functions must agree to collaborate and where appropriate, standardize on a common infrastructure. The current focus of the data centers is to create standardized data sets for national, state and local use. As envisioned, these data sets would be housed in the warehouse.

b) Storage And Management - The second major component of the framework is the Storage and Management component. This is the data warehouses component. It is the “crossroads” between data stewardship and data information usage where agreement must be reached on such critical issues as data modeling, metadata, accuracy, security, privacy, and data availability. These data warehouses function as authoritative sources of geospatial data for agency employees, other government agencies, and to the public. These data will drive e-government and program Internet applications.

Security and business continuity concerns are running high in the current national environment and these issues are critical to support agency mission critical applications

as well. Database and network and copyright security and privacy are of primary concern when making data and services available to a broad range of users. The warehouses will implement technology in accordance with USDA Electronic Access Initiative Policy to assure security and privacy of the data. In fact, the warehouse will help implement this policy by ensuring a secure method for states and local offices to disseminate their geospatial data to the public.

One of the single greatest benefits of the warehouses will be the access of integrated, enterprise databases for many users. Leveraging as much Standards-based Commercial-off-the-Shelf (SCOTS) technology and infrastructure as possible will enable this access. This warehouse architecture will leverage standards such as OpenGIS, common, and defacto standard interfaces to the greatest extent possible.

The warehouse will employ a replication strategy to provide seamless, high data availability. Each data center would be responsible for management of the primary data and a secondary source of the data would be replicated at another data center or within a web farm. This not only provides a secure backup of the data, it also provides the ability to fail over to another site should one site go down or to balance the load across sites.

There are two components of Storage and Management: the data warehouse and data marts. The warehouse is the primary source of the data. A data mart is a repository of data created from the data warehouses or other enterprise or external data sources. A data mart is designed to serve a particular community or business need. The emphasis of a data mart is to meet the decision making demands of a particular group of users in terms of analysis, content, presentation and ease-of-use. The geospatial data warehouse would reside within the data centers; however, data marts are distributed. Most data marts will reside within the Web Farms and used as the data source for Internet applications. Other data marts could be distributed to states or local use. One of the major aspects of the piloting phase of the implementation effort is to implement the technology and process for managing the replication of the data and managing the data marts.

c) Access - Access to the data warehouse is provided through seamless interfaces is the component that is seen by most of the users. This component contributes much of the business case benefits of geospatial data warehouse. Seamless, easy-to-use interfaces will contribute to a successful warehouse strategy. Users will want to add to warehouses if the data can be easily loaded and retrieved for business needs.

IMPLEMENTATION PLAN

The recommended approach to implementing geospatial data warehouse capability is through an iterative development cycle. This iterative approach is especially suited for data warehousing projects due to the evolving nature of business requirements. As the infrastructure and data model mature, more and more users accept and expect the data warehouse for day-to-day decision support. The following quotes from the Pat Garvey, Director of the Environmental Protection Agency's Envirofacts data warehouse, sums up this point

-- *“Don’t think too big, and don’t start off too grandiose. Keep expectations lowered.”*

-- *“Always plan for expansion and enhancement—expansion because you want to keep driving more data into the warehouse and enhancement because you want to be able to access the data easily and completely within the warehouse,”*

The Pilot phase is expected to last 4 months. It would implement a small-scale working model of the system, work out details, and demonstrate the feasibility of the architecture. During this phase high-risk items would be identified and mitigated and costs can be more fully fleshed out.

Phase I would follow immediately after the pilot phase with procurement of architecture items in FY 2002 and implementation in FY-2003. During Phase I critical warehouse data sources would be integrated and decision support data marts implemented. Continued effort should be made to explore other partnering data management solutions (USGS, EPA, FS).

Phase II procurement of architecture pieces would be in FY 2003 with implementation in FY-2004. This phase would complete implementation of robust architecture. Further efforts should be made to explore other storage solution partnerships.

Operations and Maintenance FY 2003 – FY 2010

The pilot will provide a roadmap for future warehouse iterations, identify and mitigate areas of high risk, and visibly show the basic benefits of warehouses. The overall implementation plan identifies the dates for a pilot and the first two iteration phases of the Geospatial Data Warehouse project. Phase II will begin the Operations and Maintenance (O&M) support for the warehouse.

SUMMARY

Business cases presented throughout this document and legislative mandates passed by Congress provide strong incentives or reasons for implementing geospatial data warehouse architecture. Authoritative, up-to-date geospatial data must be available for the partner agencies to share. The data must be easy to use and maintain, and available to the public as appropriate. The payoff of implementing the proposed warehouse strategy are substantial and far outweigh the risks of continuing to maintain loosely organized, agency-specific data collections. Several issues will need to be addressed by the partner agencies, most notably – adequate, skilled staffing to operate the warehouses; a viable network infrastructure to move data from the warehouses and marts to users or applications; and sustained budgets to support development and ongoing operations. Contract support will likely be needed to provide staffing and skills not readily available from the partner agencies. A more robust telecommunications infrastructure will ultimately be needed to implement new applications and technology that require data streaming. At this stage of planning, the cost of implementing a data warehouse architecture for the agencies is estimated at just over \$0.4 million for a Pilot, \$5.6 million for Phase I to establish the warehouse infrastructure, and \$4.7 million for Phase II to refine the infrastructure and to transition into ongoing operations and maintenance.

1. Background

The Service Center Modernization Initiative (SCMI) Information Technology Working Group (ITWG) sponsored several studies over the past year detailing geospatial data warehouse architectures and telecommunications infrastructure and the associated costs, technologies, benefits and risks for each alternative. These studies have been conducted in cooperation with the all three Service Center Agencies (SCA); Farm Service Agency (FSA); Natural Resource Conservation Service (NRCS), and Rural Development (RD), in order to determine the optimal geospatial data access and distribution architecture. This architecture includes determining the location for geospatial data warehouses, data marts and web-based geospatial applications.

The ITWG used these studies to help make procurement and funding decisions in cooperation with other related United States Department of Agriculture (USDA)-sponsored groups including Electronic Access Initiative (EAI), Common Computing Environment (CCE), and LAN/WAN/Voice (LWV). The findings presented in these studies serve as the foundation from which an implementation strategy can proceed. A brief summary of two of the studies is presented here for context.

1.1. Distributed Management of Geospatial Data

One plan, titled *Phase II: Implementation of Geospatial Data Warehouses, September 2000*¹, resulted in several broad recommendations that support the implementation of a distributed model for management of geospatial data. In order to achieve this distributed paradigm, additional recommendations were made supporting the establishment of geospatial data centers at both National Cartography and Geospatial Center (NCGC) and Aerial Photography Field Office (APFO). These recommendations increase the need to build the communications infrastructure between these two centers. Additionally, a recommendation was presented to house geospatial web applications at the web farms currently located in Kansas City, MO and Ft. Collins, CO. These recommendations were presented to the USDA Office of the Chief Information Officer (OCIO) office in December 2000. In addition to these broad recommendations, the plan detailed several distributed architecture scenarios and evaluated the investment required to implement each scenario. Investments include telecommunications infrastructure improvement, hardware, software, staff, and training.

More specifically, the plan outlined three recommendations that were presented to the OCIO:

Recommendation #1

Establish on-line data warehouses at Data Acquisition and Integration Centers (APFO & NCGC)

- Centers serve as the certified source for data dissemination and on-line applications. Centers are responsible for acquisition, integration, storage, archival, maintenance, and dissemination of geo data to internal users and

public. Centers will implement security measures (OCIO Stack) to meet department policy.

Recommendation #2

House on-line web applications at Electronic Access web farms.

- Leverage existing and future infrastructure at web farms, including high speed Internet access, robust security features, common web services, and staff support. Fort Collins web farm will provide a common point of entry (portal) into the data repositories residing at NCGC and APFO.

Recommendation #3

Establish a common Internet Portal as a “one-stop-shopping” service for geospatial data.

- The distributed nature of the data appears seamless to users by linking the warehouses through a common Internet portal that provides one-stop-shopping services.

1.2. Requirements-based Access and Distribution Models

Another study recently conducted titled, *Requirements and Cost Benefit Analysis. Managing Geospatial Data for Better Program Delivery. A Service Center Initiative²* (June 2001) outlines the business and technical requirements supporting the use of geospatial data among the Service Center Agencies. The study identifies geospatial data access and distribution architecture alternatives based on the requirements driving the need for geospatial data. Three general alternatives were presented in addition to the current architecture already in place. The alternatives include a centralized approach, a distributed approach and a mixed approach to data access and dissemination. Each alternative was weighed according to the how well the architecture met the business and technical requirements and the cost to implement and maintain using the currently implemented architecture as a baseline. In addition to the architecture alternatives proposed, the study evaluated eleven other emerging technologies that have the potential to impact the access and dissemination of geospatial data in the near future.

The result of this evaluation supports the implementation of a distributed architecture based on the number of requirements satisfied and the implied cost to implement and maintain. Additionally, the distributed architecture provides an organizationally sound approach, which is in line with current CCE architecture moving forward in Service Centers to support non-GIS related operations. Although the study supports the distributed architecture scenario, it does not make recommendations concerning the location of the distributed components or geospatial data. It is referenced here to demonstrate the support that corroborates the case for a distributed architecture model.

1.3. Warehouse Implementation Strategy and Prototyping

The recommendations put forth in these two studies support a distributed architecture for geospatial data access and dissemination. Now that a distributed architecture has been established and agreed upon, it is necessary to develop the requirements for

implementation of each of the components within the proposed architecture and identify the function each component will achieve. In addition to location and function, it is necessary to determine where the certified source of each data set will reside and how this network of data warehouses, marts, and web farms will manage access, security, and performance. These are the issues will be addressed in this implementation strategy.

Finally, this strategy will present a iterative, results-based project plan of tasks and timelines to be prototyped that can be used to measure the success and make any necessary adjustments and modifications to the strategy before implementing on an enterprise-wide scale. This prototyping will allow the department to take advantage of a tested and proven data warehouse architecture that will enhance the distribution of services to USDA customers as implementation proceeds.

2. Scope and Objective

This study will define the implementation of distributed geospatial data warehouse architecture in the context of the studies, acquisitions and implementation that have occurred over the past four years. The study will provide specific details concerning the location of geospatial data, hardware, software and security, staffing and the telecommunications infrastructure requirements that are needed to implement and maintain the operational architecture.

3. Geospatial Data Warehouses Business Case

The need to develop an implementation strategy for geospatial data warehouses stems from the programmatic developments that necessitate integrated program delivery and the advances in technology to enable on-line access and delivery of USDA services. Digital geospatial information is a central element in the delivery of services to USDA customers and has been highlighted in several high profile USDA program initiatives. There are several examples throughout government including USDA where the combination of “geo-enabling” existing business processes and on-line program delivery has demonstrated significant benefits. These benefits realized at many levels including labor savings, accuracy of information gathering, reporting and usability. Some examples include the USDA Office Information Locator, the NRCS Soils Data Viewer and the NRCS Electronic Technical Guide.

Section 204 of the new Farm Bill, Administrative Requirements For Conservation Programs^{*}, states that agricultural producers may enter into contracts with States (including State agencies and units of local government), private non-profit, community-based organizations, and educational institutions with demonstrated experience in providing technical assistance services. In addition, Section 204 states that some of the information gathered during these contracts shall not be considered to be public information and shall not be released to any person or Federal, State, local agency, or Indian tribe outside of the Department of Agriculture. This legislation has the potential

^{*} As of this report, the House and Senate have passed the bill and it is in committee now to resolve the two versions.

to effect how information, including geospatial information, and is accessed and used by non-NRCS technical assistance providers. Specifically, this mandate could require the ubiquitous availability of producer level geospatial information in a secure environment. The following are descriptions of other mandates that support the need for a USDA geospatial data warehouse initiative.

3.1. Secretary of Agriculture Food and Agriculture Policy Report

The September 19, 2001 report by Agriculture Secretary Ann M. Veneman titled *Food and Agricultural Policy: Taking Stock for the New Century*³ included a section detailing the importance of integrated programs, which detailed the importance of utilizing technological advancements such as Geographic Information Systems (GIS) to help make more informed decisions. However, in order to take advantage of these technologies, USDA needs to coordinate the availability and distribution of geospatial data and services. This cross-agency coordination will allow USDA to streamline the delivery of services to their customers.

The technological limitations of such a system, described in the Secretary's report that are currently being addressed within the agency include the ability to provide:

“one-stop shopping for delivery of services to rural America; sharing and integration of data bases and information, and computation environments across agencies and programs; and new flexibility for increased coordination of resources”.

It is this type of progress recognized at the department level that supports the need for coordinated management and delivery of geospatial data and services that the Service Center Agencies are working towards. The report describes the importance of interagency cooperation in the effort to achieve “one-stop shopping”.

3.2. E-Gov Initiative

The e-Government initiative has been designated as one of five key elements in the President's Management and Performance Plan, as outlined in the memorandum issues by the Office of Management and Budget on July 18, 2001. The focus of this initiative is centered on the following four service areas that have been identified for improvement. These are:

- *Service to individuals*: provide “one-stop shopping” for citizens.
- *Service to businesses*: reduce the burden on business through Internet protocols and removal of redundant reporting.
- *Intergovernmental affairs*: help states meet reporting requirements
- *Internal efficiency and effectiveness*: use e-business best practices to lower cost and increase productivity.

The "Geospatial One Stop" is one of 23 framework initiatives developed by the E-Government Task Force established in 2001 by Office of Management and Budget (OMB). The stated⁴ vision of the Geospatial One Stop is

"to revolutionize e-Government by providing a geographic component for use in all e-Government activities across local, state, tribal and Federal government. The implementation of the by Geospatial One-Stop in the near-term will:

- ✓ *Providing standards and models for the content of a geospatial data framework;*
- ✓ *Providing an interactive index to geospatial data holdings at the Federal and non-Federal levels; Initiate interaction between Federal, state, and local agencies about existing and planned spatial data collections; and*
- ✓ *Providing an online access point to geospatial data."*

The Environmental Protection Agency (EPA) has made a great deal of progress in the applying the E-Gov Initiative to their business model. One example of this is the EPA EnviroFacts Warehouse, which provides geospatial business information for several environmental parameters including water quality, toxic release, Superfund site location, and hazardous waste. The maps generated from the EnviroFacts EnviroMapper program can now be linked into other user applications using a feature called OpenLink. Using Hyper Text Transfer Protocol (HTTP) and XML protocols, OpenLink generates code that creates a user-defined image or map that can be inserted into a Hyper Text Markup Language (HTML) document. Alternatively, users can link directly to EnviroMapper, incorporating maps into their current application.

3.3. Security

Additional support highlighting the importance of on-line access and dissemination of geospatial data has recently been noted in a briefing by John Moller of the Federal Geographic Data Committee (FGDC), which outlines how FGDC and National Spatial Data Infrastructure (NSDI) participants can support the new Homeland Defense Working Group. The briefing emphasizes the need for shared information and collaboration and the need to link government, business and citizens through the NSDI framework.

FGDC is well positioned to address the need for access to geospatial data through the following initiatives:

- Implementation of metadata standards and procedures
- NSDI Clearinghouse and data discovery capabilities
- Data framework, which provides a portal to the most commonly used base map data sets
- Serving as a network for intergovernmental and inter-sector relationships

Salient points of the briefing that make the case for USDA data warehousing include the following:

- Shared information and collaboration are essential. Although presented in the context of homeland security, the need for access to geospatial information has been a long-standing issue promoted by FGDC.
- As more agencies register their data resources in the NSDI Clearinghouse, agencies will realize a savings in the development and maintenance of geospatial data through a “build once, use many” paradigm.

Another aspect of information access and dissemination discussed in the FGDC briefing concerns the ability to protect and maintain the physical infrastructure of the system. Weaknesses and vulnerabilities in a distributed system, whether caused by natural disaster, physical equipment failure or targeted attacks must be identified and resolved. Security can be strengthened to prevent unauthorized access to system resources through the use of firewalls, virtual private networks (VPN) and user authentication. Techniques such as mirroring, replication and load balancing can be implemented to handle system outages due to unforeseen breaches in infrastructure security.

In summary, the FGDC presentation outlined that geographic information and GIS are critical and geospatial information must be available. Ensuring the availability of this information will require that the infrastructure that provides access and delivery of information is protected and contingency plans are in place in the event of a breach in the security of the network.

4. Geospatial Data Warehouse Requirements

As one of the largest civilian federal agencies, the USDA is challenged with making informed business decisions across multiple agencies and multiple levels of government across the entire nation. The sheer number of information sources distributed across USDA makes the implementation of enterprise decision support challenging. Geography is one of the few common threads that interconnect this complex patchwork. This is why the implementation of geospatial data warehouses is so important. This strategy document will help USDA come to terms with this volume of information ensuring that it gets to the right people at the right time and in the right format.

The existence of spatial information is prolific in USDA. Whether it is data elements describing acres, zip codes, or conservation districts, USDA employees and customers are making spatially based decisions on a daily basis. The geospatial data warehousing initiative aims to make that process better, faster and cheaper by delivering certified, integrated information to the decision maker in an easy-to-use manner. The following is a list of several high-level requirements that drive the need for an integrated geospatial data warehouse. These requirements are broken down into business requirements that meet the stated business objectives and system requirements.

4.1. Business Requirements

Requirement	Description
<p>Maintain an infrastructure that provides the most up-to-date USDA certified source for geospatial data.</p>	<p>Data warehouses should provide an organized and efficient infrastructure for centralized management of certified source, national data sets. Certified source data is the official data set of an agency whose mission includes the production, maintenance, update and dissemination of the original data and replicated subsets. Dissemination includes providing on-line access as well as other data delivery methods.</p> <p>Additional accommodation should allow for access and integration of data sets that USDA is not the certified source provider, such as commercial value-added data resellers; on-line data service providers; and state and local governments that have access to more accurate and regionalized data sets. This infrastructure allows users to select from the best available data for their particular application and needs.</p>
<p>Provide on-line access to national data sets to meet emerging program and E-Gov requirements.</p>	<p>Provide on-line access to data for meeting program needs and emerging needs of farm bill and E-Gov requirements. Provide databases that can be linked to on-line business applications and accommodate on-line data viewing.</p>
<p>Provide high availability of data with reliability and integrity of data intact.</p>	<p>Users of the data warehouse will expect a high degree of availability and reliability to support their 24 x 7 x 365, anywhere operational requirements. The data warehouse components must support load balancing, fail-over, disaster recovery, and data replication in order to meet user demands.</p>
<p>Provide a mechanism for State and local technical information to be provided to the public and the USDA enterprise</p>	<p>The management of local data presents a challenge for data management. As more and more applications are implemented to operate over the Internet, these applications could benefit from and in some instances require state and local data. For example, the new Farm Bill provides for private organizations to assist in developing conservation plans. These private organizations will be certified by the NRCS and will use NRCS technical information that is available in the local Field Office Technical Guides. To facilitate this process a major project is under way allow access to Technical Guides through the Internet. This will require in some instances state and local data be provided to the public. USDA security policy does not allow state or local systems to be accessed by the public unless extensive and expensive security measures are implemented which in almost all instances are cost prohibitive. The public will not be allowed to access state or local servers. For the data to be public the server will need to be part of the network where it can be accessed by servers that are in the web farms or data centers or the data will need to be replicated at in the web farm or data center.</p>
<p>Business Continuance</p>	<p>The system must support the ability to continue business during prolonged system outages or impeded system performance. This includes times during system failure, system maintenance and upgrades. This can be accomplished by replicating those data that have a high business dependency.</p>
<p>Support privacy and security requirements.</p>	<p>Warehousing must be able to allow or restrict access to specific data sets by specific user groups. Risks that need to be addressed include: Data copyrights – USDA has purchased data for limited use as background data. This data can be used for viewing only and is not to be freely distributed for download.</p>

	<p>Database corruption – Care must be taken to ensure that access to certified source data is protected.</p> <p>Data aggregation – Data warehousing applications provide an opportunity to aggregate and analyze data in ways that were previously not available. Care must be taken to assess each data element in this context and provide a mechanism that protects against this misuse of data while still provide this customer service to appropriate users of this information.</p>
Support metrics for system administration.	Program managers, data stewards and data managers must have access to metrics that describe the data in the warehouse and the access and usage of the data. These metrics should describe the volume of data processed, who accessed what data when, the version of data retrieved, size of data sets delivered, and method of delivery (CD ROM, FTP, etc.) and the overall system health, configuration and content.
Provide feedback mechanism for users.	There must be a mechanism for users to ask for assistance and gain feedback from data stewards and the system administrators. There must be a mechanism in place to track requests for data as the requests are being processed.
Support delivery in multiple formats as requested by user (within limits, structure remains unchanged).	Just as the data warehouse ingests data from a variety of sources, it must support the transformation to a variety of information consumers. The transformations can take many forms including, binary transformation, projection transformation, generalization, aggregation, interpolation and interpretation. These transformations must be automated and repeatable to the greatest extent possible.
Ability to incorporate full support of metadata.	The system must have the ability to manage and support metadata. Metadata is key to supporting the search, discovery, and analysis of data from the warehouse as well as the development of applications that use warehoused information. Metadata must be tightly coupled with the data it describes and not managed in an altogether separate environment that risks a synchronicity.
Support procedures and data update of data marts.	As specific data subsets (data marts) are deployed throughout the enterprise including national, state, and local level organizations, the data warehouse must allow for consistent, transparent, and repeatable replication of data from one node to another. Common framework specifications and data content models would govern the deployment of such replicate data marts.
Facilitate acquisition, integration and distribution of warehouse data.	The data warehouse must accommodate the automation of information flow to the greatest extent possible. This means that the flow of information from operational data stores, through the warehouse and to the data marts and applications must be well understood and controlled through well-known interfaces so as to maintain the integrity of the data as it passes through the value chain.
Support versioning configuration for on-line and off-line migration.	The temporality of information is important to applications such as resource change detection, ownership tracking and planting history. The data warehouse must respond to the decision support requirements for on-line history tracking and off-line archival of information once it has passed its current usefulness.
System should be user friendly, but meet the needs of more knowledgeable customers and support custom application development.	The geospatial data warehouse, when operational will be a complex and interrelated set of components. However, that complexity should not manifest itself to the users or developers of the system. An iterative approach should be used to deploy the data warehouse so as not to paralyze the very business operations that are supposed to be enhanced. Conversely, the system must be robust enough to be useful to the most sophisticated users.

Address user requirement to deliver integrated data in specific geographical regions.	One of the greatest benefits of the data warehouse is the ability to rapidly and repeatedly deploy integrated warehouse information into specialized data marts that can be used throughout the geographically diverse and dispersed USDA enterprise. The data warehouse must be able to respond to an ever-changing demand for integrated business and geographic information.
--	---

4.2. System Requirements

Requirement	Description
Framework for providing common services (objects, COM, Net tech, etc.).	System must have the ability to support common data services. To minimize lifecycle time and effort, a common framework for developing applications that use warehoused information must be developed and communicated to developers.
Support OGC standards, currently and future. Including the ability to integrate multiple geodatabases in multiple data stores.	Open interoperability standards such as those from the OpenGIS Consortium (OGC) provide framework architectural specifications from which operational systems can be implemented. As applicable standards come to fruition by means of standards based commercial-off-the-shelf (SCOTS) software, USDA reduces the risk of obsolescence by implementing proprietary solutions.
Integrate data to Service Center (SC) standards.	The data warehouse should support standardization where needed, but more importantly support flexibility to the greatest extent possible. This will ensure a far-reaching integration of the data warehouse into currently disparate systems.
Provide security measures for internal and external data sources	NRCS has cooperative agreements with outside organizations to house NRCS data. When the USDA OCIO implemented firewall measures for USDA centers, it became apparent that similar measures would have to be implemented at these offsite locations or the data would have to be moved to a USDA center having the OCIO security stack.

4.3. Geospatial Data Warehouse Vision (Technical Architecture)

Section 5.4 of the initial study entitled *Phase II: Implementation of Geospatial Data Warehouses* described the vision for the geospatial data warehouse. The description contrasted the industry wide understanding of a data warehouse with the USDA notion of a geospatial data warehouse. In this document, certain semantics of that notion are clarified effectively bringing these notions closer together.

4.4. Building a Data Warehouse for the “Vision”

In order to migrate from the loosely managed USDA geospatial data environment of today to a coherent and well-defined data management environment, a common vision for USDA geospatial data warehouses must be developed. According to Inmon, in "*What is a Data Warehouse?*"⁵, a **data warehouse** is:

- **Subject-Oriented** - Oriented around the major subjects of the enterprise. These subject areas can be data oriented (e.g. soils, farms, demographics) or process/function oriented (e.g. conservation planning, crop reporting, lending).
- **Integrated** - Data found within the data warehouse has consistent naming conventions, consistent measurement of variables, consistent encoding structures,

consistent physical attributes of data, etc. For example, two separate operational systems may store land area as acres and the hectares respectively. In the data warehouse, the values would be converted and stored as one or the other.

- **Time-Variant** - All data in the data warehouse is accurate as of some moment in time. For example, the tract of land may represent the ownership boundaries from the time of purchase to the time of sale.
- **Nonvolatile** - There are only two kinds of operations that occur in the data warehouse - the initial loading of data, and the access of data. There is no update of data (in the general sense of update) in the data warehouse as a normal part of processing.

In contrast a **data mart** is a repository of data created from the data warehouse or other enterprise or external data sources. A data mart is designed to serve a particular community or business unit. In scope, the data may be derived from an **enterprise-wide database** or **data warehouse**. The emphasis of a data mart is to meet the decision making demands of a particular group of users in terms of analysis, content, presentation, and ease-of-use. Users of a data mart can expect to have data presented in terms that are familiar. Web enabled data marts and user specific portals are becoming a dominant influence in the presentation of data marts.

The USDA Data Management Team adopted these definitions and implemented a small number of limited data warehouses and data marts prototypes. Considering the business requirements described earlier, the traditional definitions of data warehouses and data marts make sense for geospatial information. To more clearly understand how a geospatial data warehouse fits the traditional notions, one must understand the components of the current geospatial data management environment. Figure 4-1 illustrates that the current geospatial data management environment is a loose collection of subject oriented spatial flat files, metadata, and tabular data. In most cases, the spatial data is integrated vertically (i.e., data is spatially referenced), but is loosely **integrated** with attributes from other functional business data. Data integration generally occurs in disparate stand-alone applications on the users' desktop. As for **time-variance** and **volatility**, much of the data represents a "snapshot" of geographic space in a given time period; however, wholesale updates are typically performed rather than transactional updates. Very little geospatial information is stored for historical reference. Thus, the current environment could be characterized as a collection of loosely integrated, stovepipe systems.

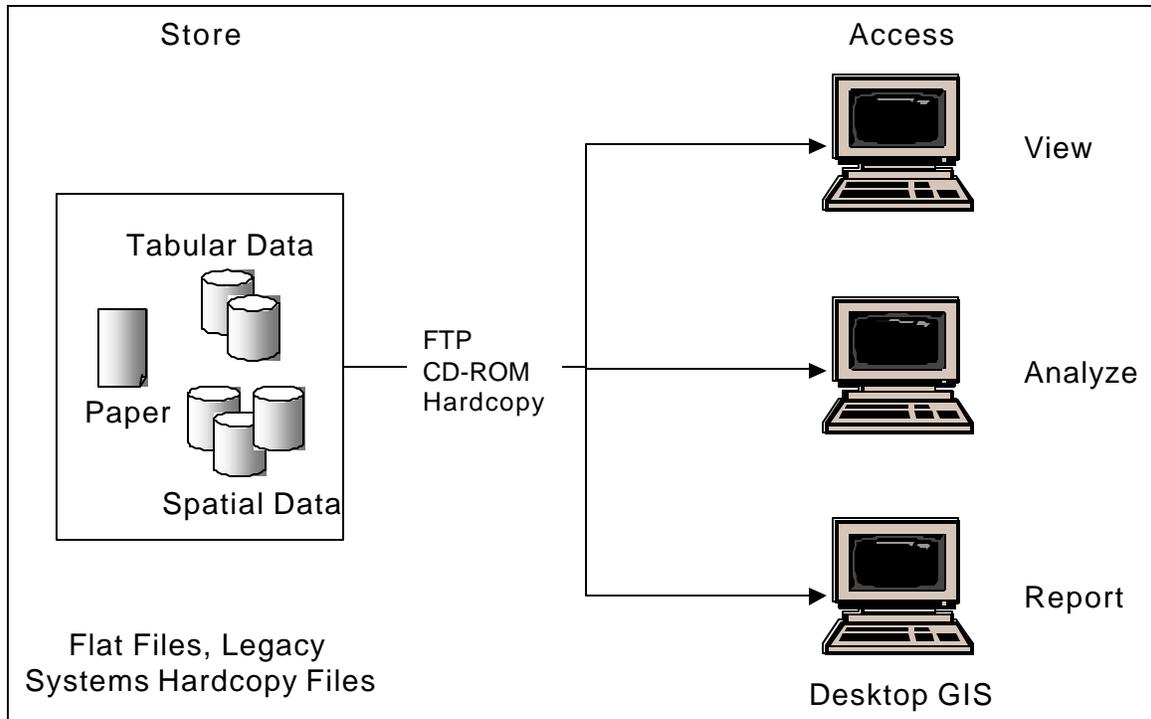


Figure 4-1 Current Geospatial Data Management Environment

4.5. Architecture Alternatives

There are several factors to consider when determining the most appropriate Geospatial Data Warehouse architecture to implement. Orr¹⁰ defines three alternatives data warehousing architectures in the following description.

“Virtual or Point-to-Point Data Warehouses

A virtual or point-to-point data warehousing strategy means that end-users are allowed to get at operational databases directly using whatever tools are enabled to the "data access network." This approach provides the ultimate in flexibility as well as the minimum amount of redundant data that must be loaded and maintained. This approach can also put the largest unplanned query load on operational systems.

As we will see, virtual warehousing is often an initial strategy in organizations where there is a broad but largely undefined need to get at operational data from a relatively large class of end-users and where the likely frequency of requests is low. Virtual data warehouses often provide a starting point for organizations to learn what end-users are really looking for.

Central Data Warehouses

Central Data Warehouses are what most people think of when they first are introduced to the concept of data warehouse. The central data warehouse is a single physical database

that contains all of the data for a specific functional area, department, division, or enterprise. Central Data Warehouses are often selected where there is a common need for informational data and there are large numbers of end-users already connected to a central computer or network. A Central Data Warehouse may contain data for any specific period of time. Usually, Central Data Warehouses contain data from multiple operational systems.

Central Data Warehouses are real. The data stored in the data warehouse is accessible from one place and must be loaded and maintained on a regular basis. Normally, data warehouses are built around advanced RDBMs or some form of multi-dimensional informational database server.

Distributed Data Warehouses

Distributed Data Warehouses are just what their name implies. They are data warehouses in which the certain components of the data warehouse are distributed across a number of different physical databases. Increasingly large organizations are pushing decision-making down to lower and lower levels of the organization and in turn, pushing the data needed for decision making down (or out) to the LAN or local computer serving the local decision-maker.

Distributed Data Warehouses usually involve the most redundant data and, as a consequence, most complex loading and updating processes.

The distributed Geospatial Data Warehouse architecture was recommended by the Data Management group in the previous study and is put forth in this document. As pointed out in the excerpt above, USDA is pushing the decision-making down to lower and lower levels of the organization. In fact, much of the valuable data for decision-making is distributed among local and state organizations. So, the need for a distributed mechanism for accessing enterprise-wide data is an inherent requirement of the Geospatial Data Warehouse. However, This concept should not lead to the conclusion that all data (local, state, national, internal, external) should be moved to a single, national “warehouse” server farm. On the contrary, the architecture should optimize the use of distributed computing resources to the greatest extent possible and rely on common interfaces to support the searching, finding and access of geospatial data resources. In the end, national servers should only warehouse data that is a national asset (i.e., soils, common land unit) or provide a service to other entities where hosting data and/or service are inappropriate (i.e., funding, security, infrastructure). Other entities should provide their own data and services that fit into the interoperability model adopted by the USDA Geospatial Data Warehouse.

With that said, there is always a trade-off to be made when deciding where the “warehouse” and “mart” data should reside. As per the definition, data should be integrated around subject areas. This approach tends to favor centralization. Conversely, since much of the data is already stored locally, it seems to make more sense to host the data near the creation point. A rule-of-thumb is to assess what value is being added to the

usability of data by warehousing. A familiar analogy can be made using a point-of-sale data-warehousing model for a grocery store chain. One of early uses of data warehousing technology was to access the cash register transactions in stores to more effectively provide products and services based on customer behavior. To gain a broader perspective on what was happening at the point-of-sale, the data was collected locally, transmitted to a central warehouse, integrated and deployed to decision support systems via data marts. The decision to integrate at a location other than the local store is driven by the need to integrate enterprise-wide data. The information that comes back to the local chain may be a value added report that drives the inventory decisions of the stores. The information about what products belong in certain shelves may be only useful to that local store and, therefore, should not be centralized.

The same is true for USDA geospatial information. Where much of valuable information is local, integration with data outside the local office provides added value. The following is a list of added values of the geospatial data warehouse.

Added Value	Benefit
Repeatable, uniform integration	Much of the time spent by USDA GIS field specialists or other employees is spent manipulating geospatial and tabular data to make it usable for analysis. Uniform processes established outside the local office provide economies of scale to the process.
Distribution to non-local users	As geospatial data becomes integrated with other geospatial and tabular information through the enterprise, the value and interest in the information will likely rise. Local offices are not the best dissemination point for information to the public, partners and policy makers. That infrastructure is more appropriate for a state or national level web farms.
Security	Current USDA policy does not permit public access to data other than through the national web farms. The data warehouse/data mart model could help coordinate this data management issue.
Data Replication	USDA has made significant investments in the acquisition and integration of geospatial data. Where appropriate, this data should be protected from compromise by replicating it to one or more alternative locations. This requirement is made somewhat more manageable by first consolidating the information. An

	additional benefit to replication is the load-balancing and fail-over functionality that comes as a possible bi-product to replication.
--	---

If no value is added to the data and the information is usable and accessible from local and external sources, then there is no need to move data beyond those sources. Definition and implementation of standard interoperable interfaces (e.g. data or web services) should be established to make the mechanics of access transparent to the end user. Environmental Systems Research Institute (ESRI) shape, image and ArcXML (AXL) formats are likely defacto for an ESRI-based enterprise such as USDA. The OpenGIS consortium defines a set of Open Standards for interoperability. ESRI currently builds Standards-Based Commercial-off-the-Shelf (SCOTS) software those implements OGC standards.

Figure 4-2 depicts a first cut at the organization of national level assets within the data warehouse architecture. The data stores to the left represent the internal and external operational data stores at all levels of the enterprise. The center section represents the collective integrated data warehouse. Layers in red indicate primary data sources, layer in blue indicate secondary (i.e., replicated) data sources. Data Marts (shown in yellow) should be located to best meet the requirements of the data users. Data Mart applications are shown to the far right. An overarching data and service catalog is crucial to the success of the data warehouse operation.

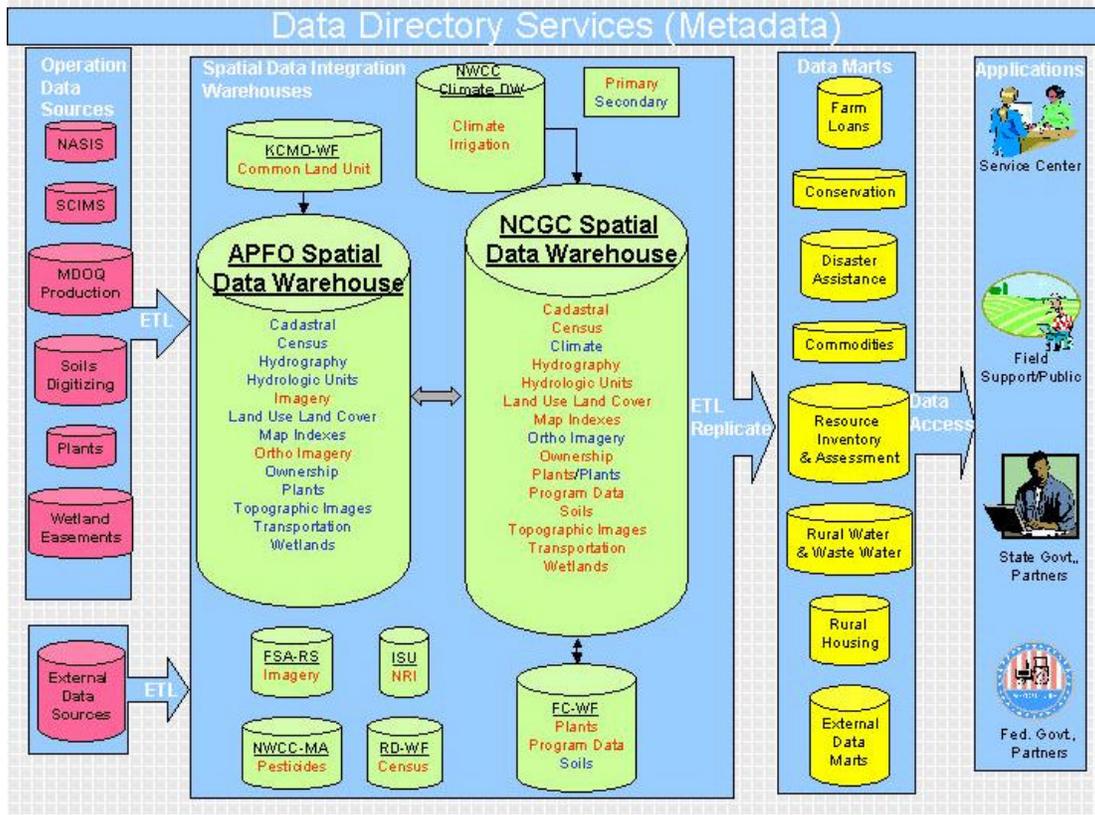


Figure 4-2 USDA Geospatial Data Warehouse Architecture Example

Figure 4-3 depicts the major components and information flow through the USDA distributed, integrated Geospatial Data Warehouse. The three major components (Acquisition and Integration, Store and Manage, Access) and their sub-components are defined as follows:

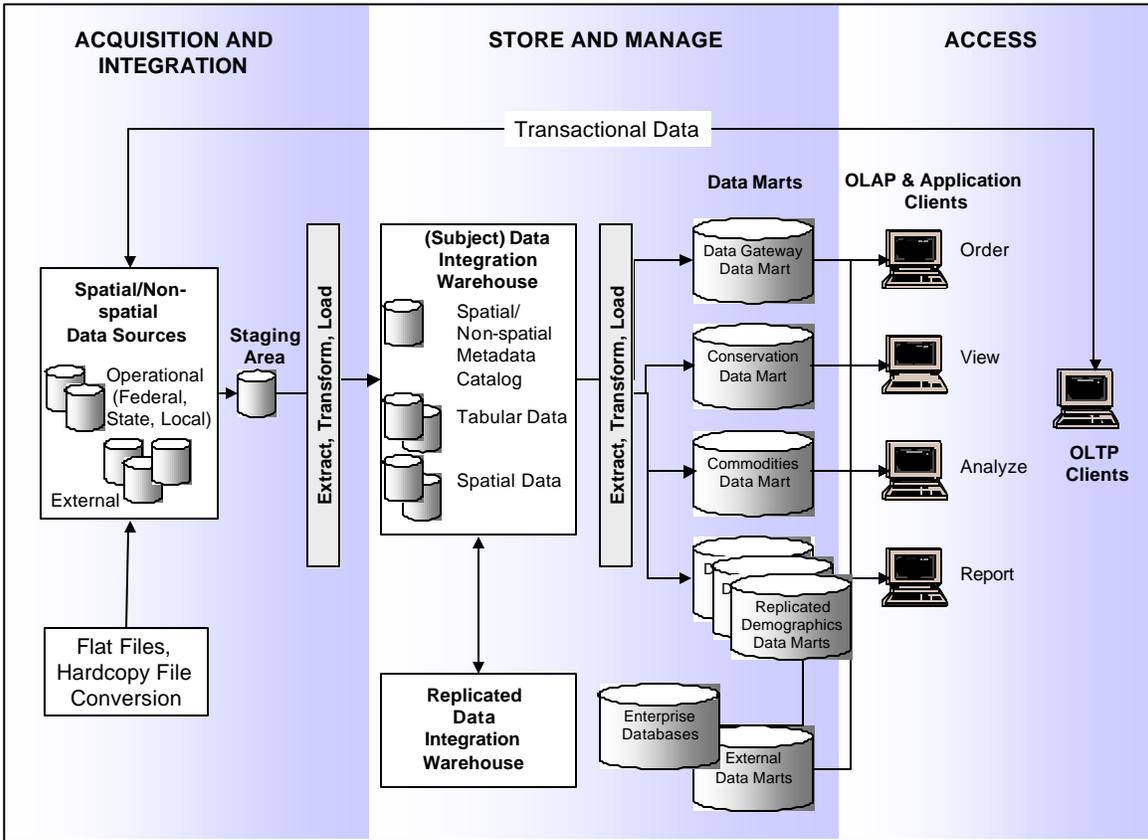
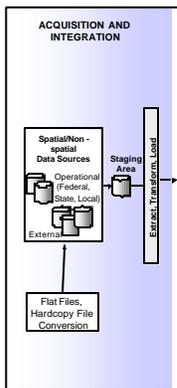


Figure 4-3 USDA Abstract Data Warehousing Architecture

4.6. Acquisition and Integration

The first major component of the Geospatial Data Warehouse initiative is the Acquisition and Integration. Much of this function currently exists or is emerging within the enterprise. For example, a large portion of the work conducted in existing USDA spatial data centers involves the acquisition/integration of spatial data from internal processes or external providers. Likewise, the current farm records and National Soils Information System (NASIS) are essentially legacy systems that could potentially feed the future data warehouse. To leverage the functionality of the geospatial data warehouse, owners of these systems must agree to collaborate and, where appropriate, standardize a common infrastructure. Of primary importance are the components of Staging and Extract, Transformation and Loading (ETL), critical to populating the warehouse. The following sections describe each component in detail.



4.6.1. Operational Data Sources

The most valued asset of any data warehouse enterprise operational data. The term operational data is most often used to refer to data that is generated by an on-line transaction processing (OLTP) system, but should include any data that is used to maintain daily operations of USDA's business. This includes internal and external data that is used for reference or analysis. Operational data will be acquired, integrated, cleansed, translated and ultimately used to populate the data warehouse. In many cases, this data is maintained by local, state or national entities within the USDA enterprise.

4.6.2. External Data Sources

Examples of external data providers that could potentially feed the data warehouse are:

- Federal partners (e.g. USGS, Census Bureau, etc.)
- State partners (e.g. Dept. of Natural Resources, Dept. of Planning, etc.)
- Local partners (e.g. county assessors office, conservation districts, etc.)
- Commercial sources (e.g. satellite imagery providers, road network providers, etc.)

A considerable amount of labor is currently invested in the manipulation of external data sources in preparation for internal use. A great deal of geospatial data standardization progress has been made as a result. However, there is a significant amount of standardization and automation to be done prior to populating the warehouse with external data sources.

4.6.3. Staging Area

Operational systems such as OLTPs or geospatial data production centers are not necessarily good platforms for staging data prior to loading into the warehouse. In some cases, the data must be co-located onto a single platform so that complicated transformation routines and quality control functions are not hampered by operational constraints. Additionally, over burdening existing production systems with routine ETL processes would be counter-productive. Several staging areas may be needed depending on the nature of the data and the intended data warehouse target.

4.6.4. Extract, Transform, Load (ETL)

ETL will likely be one of the most complicated and costly components of the data warehousing effort in terms of software, management, coordination and processing. In most warehousing efforts, the cost of extracting, cleaning, and integrating data represents 60-80% of the total cost of a typical data-warehousing project.⁶ This may not be as severe for USDA geospatial databases, but will still be significant. ETL is a process performed to integrate and transform data from an operational data model to the model of the data warehouse. Specific processes typically includes:

- Data renormalization
- Data aggregation
- Standardization of data elements
- Elimination of duplicate data elements

ETL can be a difficult task typically hampered by poor data quality. Problems include data fragmentation, redundancy, accessibility and completeness. To illustrate this problem, consider a layer often used within USDA to navigate to a customer's property, the Public Land Survey System (PLSS). Each record in the PLSS database includes a column that stores the concatenated representation for a section that includes: Section Number, Township Number, Township Direction, Range Number, and Range Direction. Different producers typically represent this key field differently based on the data generator's specific needs. A similar type of problem can extend into the spatial representation of the data as shown by the sliver anomaly in Figure 4-4.

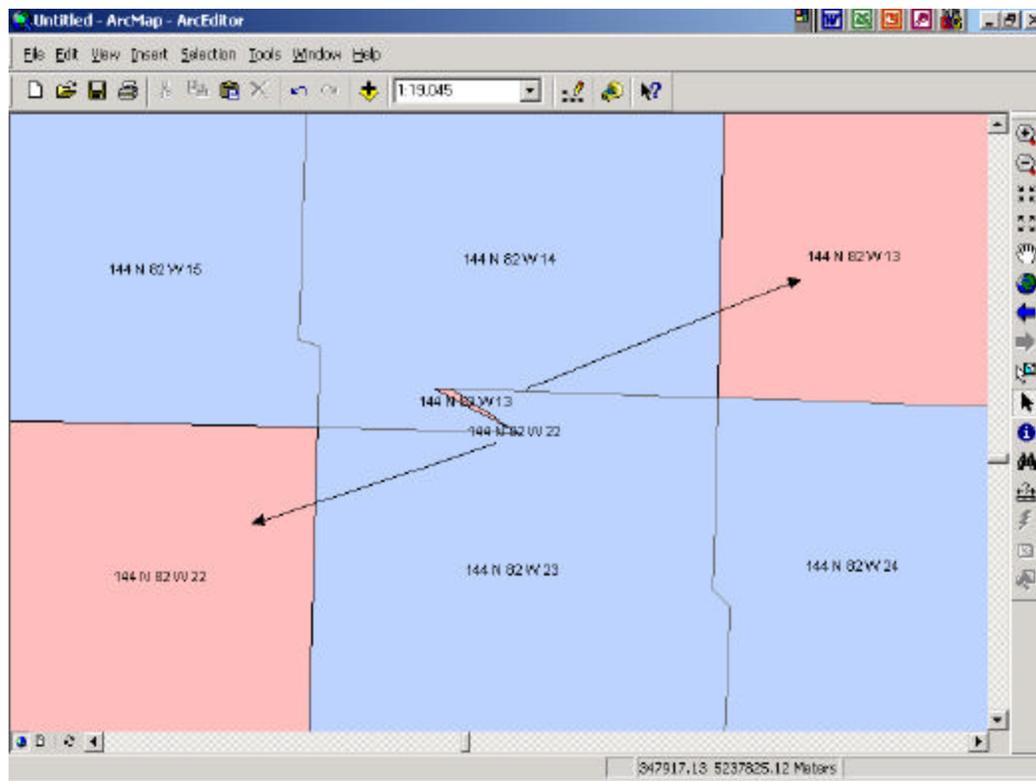
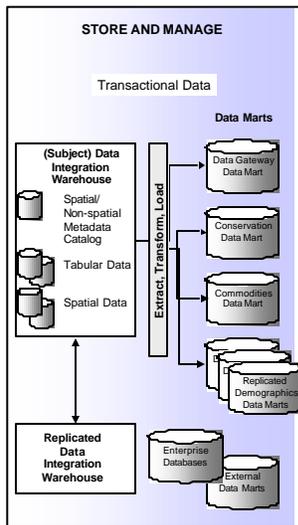


Figure 4-4 PLSS Data Anomalies

Inconsistencies and anomalies such as these are being fixed in an ad hoc fashion but could become part of the ETL process executed each time the warehouse is refreshed. This process will tax the resources of those on the integration staff if not done in a programmatic fashion.

4.7. Store and Manage

The second major component of the Geospatial Data Warehouse is the *Store and Manage* component. This portion of the overall data warehouse includes both certified source data warehouse databases as well as a wide variety of decision support data marts. Round-the-clock data availability is another critical aspect to the store and manage component. Decision support systems that feed critical business functions must be supported by an enterprise storage plan that supports data redundancy, mirroring and business continuance.



4.7.1. Metadata Catalog

Metadata, is not only critical to the spatial Geospatial Data Warehouse storage and management component, but is the thread that weaves throughout the data warehousing system. Metadata facilitates the understanding of the information model from both the technical and business points of view. Metadata describes the structure of the information, the quantitative and qualitative nature of the data content, as well as the services offered by the data providers. Metadata management in a large, widely distributed environment such as USDA is a challenge for many reasons. For example, much of the data at hand comes from disparate sources whose policies and practices for collecting and maintaining metadata are varied and inconsistent. Additionally the tools for maintaining and exchanging metadata from the data sources vary widely. National efforts to standardize the content and distribution of geospatial metadata by the Federal Geographic Data Committee (FGDC) provide some level of consistency in the spatial domain.

Management of metadata within the Geospatial Data Warehouse has two primary focal points.⁷ 1) From the technical view, data managers, database administrators, and developers must understand the constructs and meaning of the information model in order to make decisions regarding the processing of data. The tools needed for this task allow users access to the underlying data model of the data to be warehoused as well as the warehouse and data marts themselves. 2) From the business perspective, the tools for analyzing the data structure is not nearly as important as the tools for discovery of data and defining the meaning and quality of the data content. The metadata catalog becomes the focal point or hub for these processes. Since the data warehouse becomes the certified source for decision support, the metadata describing it must also maintain similar credibility. This means that the metadata catalog must:

- Be synchronized with the data that it describes
- Provide a simple interface to reflect updates to the data warehouse
- Allow for the notification or discovery of new data as it is loaded to the data warehouse
- Provide adequate description of the heritage of data from its native source to each decision support system that uses it

4.7.2. Tabular Data

There are many data warehouse reference implementations for tabular data. USDA is conducting a small number of data warehousing projects within the Service Center Agencies. What distinguishes a tabular data warehouse from a geospatial data warehouse is that all information is stored in simple data types (i.e., characters, integers, etc.) and can be manipulated using specialized data warehousing DBMS tools. Much of the spatial data within USDA is attributed with tabular data that describe the properties and characteristics of the spatial objects. While it is possible to store attributes of spatial data along with the geometries that define the location and shape of those objects, many legacy data sets within USDA still maintain spatial information separately from their tabular attributes. This is often the case for more complex data models that have a series of related look-up tables that link codes and abbreviations to more descriptive meanings. Figure 4-5 illustrates this point. A Soil Survey Geographic Database (SSURGO) map unit consists of between one and three components; each component has 60 properties and interpretations that captured in 84 different data elements (attributes); each component will have between one and six soil layers; each soil layer may have up to 28 properties.

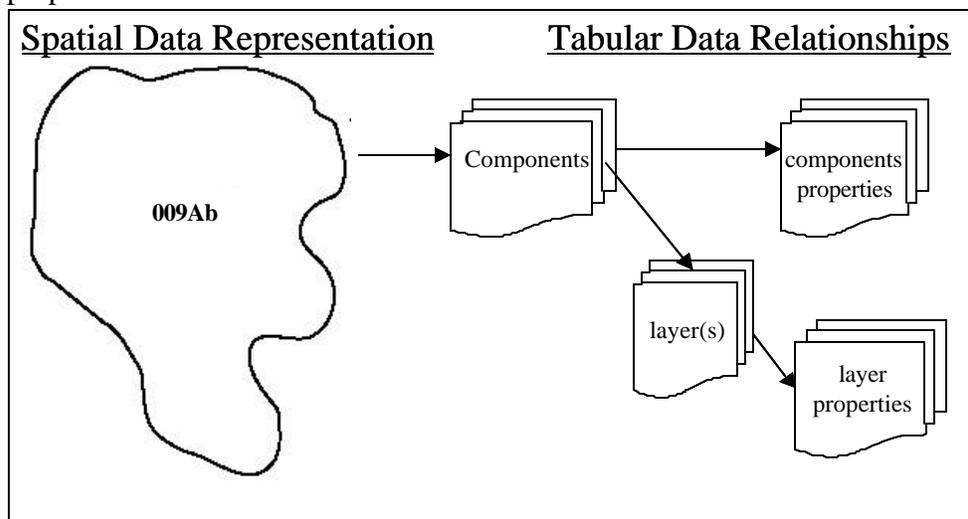


Figure 4-5 SSURGO Tabular Data Structure Diagram

This figure is a simplification of the complicated structure of one-to-many relationships between map units, components and soil properties. Related tables are associated with one another through a composite join that consists of a map unit identifier and a sequential number, a unique identifier for each component. The tables are related to the spatial data through the map unit identifier.

4.7.3. Spatial Data

From the description in the preceding paragraph, it is clear that tabular data is a key component to the spatial information model. However, spatial data has some characteristics that sharply contrast with typical tabular data. These are:

- Spatial data elements are typically modeled with specialized constructs that are just now appear as part of the mainstream Relational Database Management System (RDBMS) product software. For example, a geographic entity might be represented as vector, raster or gridded entities that require special handling in the database. Spatial handling tools are not widely available in the data warehousing industry.
- Depending on the amount of detail represented in the spatial database, the file sizes can be quite voluminous compared to typical tabular data.
- Portrayal of entities is often abstracted to a resolution commensurate with scale. This can be problematic when aggregating information through multiple levels of detail.

4.7.4. Replicated Data Integration Warehouse

Data replication is an ambiguous term used to describe a family of technology. Data replication satisfies multiple technical requirements depending on how it is deployed. This section will cover the replication components required to satisfy the business requirements for spatial data decision support. The various reasons that data should be replicated includes the following:

- Business Continuance:
 - Real or near real-time version of the data must be stored at a location that can be used to “cut over to” in the event of a catastrophic event to the primary data warehouse.
 - Minimize the impact of warehouse reloads by maintaining a production warehouse and a staging warehouse and cutting over to the staging warehouse once it has been re-loaded.
- Performance:
 - Provide multiple data warehouse or data mart nodes that can be used in a load-balancing scheme. When one copy is busy, the other can seamlessly take over.
- Database synchronization:
 - When migrating data from the data warehouse to the various data marts, database replication is a technology that is commonly used to move large blocks of data from one store to another without having the overhead of query.

Business continuance and performance data replication would likely employ mirroring replication (exact replicas of data), whereas database synchronization might employ a subscription type of replication (subsets of the data). Again, there are several approaches that may be used to satisfy this requirement. Details of various replication technologies will be described in Section 8.1.1.2.

4.7.5. Extract, Transform, Load

Section 4.6.4 described the extraction transformation and loading of data into the data warehouse. This section describes the ETL processes that are required to load the on-line analytical processing (OLAP) data marts. The primary characteristic of the data mart ETL process is the to populate customized data marts that support departmental Decision Support Systems (DSS). Common customizations⁸ include:

- **Subsets** – Selecting specific subsets of data in the data warehouse for a particular application or analysis. For example, all of the fields enrolled in the 1998 Conservation Reserve Program (CRP) sign-up year.
- **Aggregations** – There are different approaches that may apply to different data being summarized, to different ways in which the aggregated results are calculated, or to different sets of categories by which the aggregated data is organized. Lending programs may summarize their data one way, while housing programs will summarize theirs another way.
- **Supersets** - One department will de-normalize their OLAP data by joining data from tables A and B, while another department will join data from tables B and C. For example, crop insurance analysts may join soils and policy data into a data mart, while CRP might join soils and the CRP sign up databases.
- **Indexing** - One department will index their data on keys ABC and BCD, while another department will index the same data on keys CDE and DEF, and so forth, to provide more optimal search paths that meet their different departmental requirements for informational processing.
- **Derivations** – A department may want a particular metric pre-computed and the results stored in their OLAP environment, such as the Revised Universal Soil Loss Equation (RUSLE) or Highly Erodible Land (HEL). A similar metric may be stored at the organizationally structured level, but the department wants to compare their department-specific calculation to the organization-standard one
- **Arrays** – In order to make the data in their OLAP environment more useful, a department may opt to create an array of data to assist them in their informational goals. For example, data that is stored one record per month in the organizationally structured detail may be required as an array of 13 months to represent a contiguous year and facilitate current-year-previous-month analysis. A spatial time-series that shows historical plantings is an example of this in the spatial domain.

4.7.6. USDA Data Marts

There is a significant need for spatially enabled decision support tools within USDA. Distributed, business oriented data marts are a critical piece of the decision support infrastructure. There exists today, a handful of spatially enabled data mart “like” implementations (i.e., Resource Data Gateway, Office Information Profile) throughout the USDA enterprise. However, they are not necessarily part of a formal warehousing infrastructure.

The implication of spatial enabled data marts within USDA is significant. As the data warehouse emerges, the need for customized, decentralized views of information and the need to migrate away from legacy systems (including spatial) will grow⁹. This, in turn, will drive the size and capability of the data warehouse to feed such marts. Forecasting the demand for USDA data marts is difficult and will depend largely on the adoption by business units as familiarity grows. However, the infrastructure to support the creation and maintenance of data marts must support key technical requirements such as loading, refresh, replication and metadata management.

4.7.7. External Data Marts

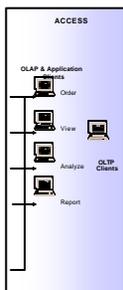
USDA currently depends heavily on external information to support business processes. This is especially true for geospatial processing where much of the spatial data is under the stewardship of outside organizations. Operational use of data directly from external sources is a common practice in USDA. However, if that same data is used across many parts of the enterprise then the overall management of that data becomes inefficient. Inmon writes that:

“When external data is fed directly into the OLAP (operational) environment, the implication is that there is no other corporate use of it outside of the department that controls that OLAP instance when there is a corporate need for that external data, then the data is fed into the organizationally structured portion of the data warehouse where it is then available to any instance in the OLAP environment”⁸

An important development in this area is the emergence of “web services” technology that is being adopted by popular vendors such as IBM and Microsoft. Web services provide the means to encapsulate business processes, publish them as services, search for and subscribe to other services, and exchange information throughout and beyond the enterprise. Web services are a major influencing technology in Internet-based geospatial technology through the OpenGIS and GIS vendor community at large. Details of this will be discussed further in section 7.2 covering the role of OpenGIS.

4.8. Access

Application access is the most visible part of the data warehouse. Therefore, this component contributes much of the benefit to the business case of the geospatial data warehouse. The success of the warehouse decision support tools is driven by the ease-of-use and the benefit that can be derived from the information. It is critical to keep in mind that the more successful a data warehouse strategy is, the more end-users are going to want to add to it.¹⁰



4.8.1. OLAP & Application Clients

OLAP and application clients will provide a much-needed benefit to the business community within USDA, their customers, partners and overseers. The sophistication of those tools varies widely from simple map backdrop applications to powerful geospatial analysis tools. The current selection of

ESRI (Environmental Systems Research Institute, Inc.) as a GIS vendor and the potential adoption of OpenGIS standards will offer robust possibilities. For example, web enabled map services could be provided as data marts serving a specific community of users. Those data marts, fed by the Geospatial Data Warehouse, could be cataloged and discovered using current technology to facilitate the sharing of common information nodes throughout the enterprise.

4.8.2. OLTP Clients

To this point in the document, the benefit of the data warehouse has been discussed in the context of decision support (analysis and reporting). The data warehouse and data mart concept is equally beneficial to OLTP type operations. Many of USDA's digital business transactions occur with very little or no context of digital geospatial information. The availability of geospatial information provided in an on-line warehousing environment would greatly enhance the usability, accuracy and timeliness of transactions throughout the enterprise.

4.8.3. Geospatial Data Warehouse Architecture Vision Summary

In summary, the key characteristics of this architecture are:

- Metadata is carefully cataloged and made available to users, administrators and developers throughout the data warehouse architecture.
- Information is loaded from operational and external systems on an as-needed frequency determined by the requirement of the data mart users.
- Spatial data warehouses are distributed to primary locations defined mainly by data ownership. However, some of the data is replicated to secondary locations.

5. Current Data Center Architecture

The implementation of the Geospatial Data Warehouse architecture defined in this plan relies heavily on leveraging existing resources. Each of the geospatial data centers serve as geospatial service and data providers for their respective agencies. Data centers roles will expand as they are integrated into the overall framework for Geospatial Data Warehousing. In addition to maintaining their current role as a service/data provider, these centers will interact more with each other, as the push for "one-stop-shop" and "seamless" services are implemented. This interaction will require each of the data centers to upgrade and/or enhance their existing infrastructure, supporting technologies and the staff that maintain these systems.

This section briefly identifies the data centers that play an integral part in the Geospatial Data Warehouse architecture described in this plan. The current state of each data center is examined in more detail in **Appendix A**. The goal of this section and the appendix is to identify the current role of the data center, their existing architecture and the resources they need to acquire in order to move forward with the additional responsibilities put upon them as members of the broader geospatial data warehouse architecture.

5.1. APFO

The Aerial Photography Field Office (APFO), Farm Service Agency (FSA) is the primary source of aerial imagery for USDA. APFO is currently engaged in several geospatial support activities for the Service Center Agencies such as MDOQ (Mosaicked Digital Orthophoto Quadrangle) production, support to the National Resources Inventory (NRI) and contracting management of the Common Land Unit (CLU) digitizing effort.

5.2. NCGC – Fort Worth, TX

The National Cartography and Geospatial Center (NCGC), NRCS is a technical leader in cartography, remote sensing, global positioning systems, and GIS. NCGC is a clearinghouse for geospatial data sets as well as a center for the preparation of maps and map products. NCGC provides innovative leadership in the map-making process. Generally, NCGC supports the soil survey process by providing map products and materials necessary for the completion of soil survey publications.¹¹

5.3. NRCS ITC – Fort Collins, CO

The Information Technology Center (ITC), NRCS is the IT development center and web farm for the NRCS. Many of applications used by Service Center Personnel are developed and supported from ITC including the Soils Data Viewer, Customer Service Toolkit, and the Resource Data Gateway.

5.4. NWCC – Oregon

The National Water and Climate Center (NWCC), has the lead in NRCS to acquire, develop, and transfer water and climate information and technology to the agency and its partners to support natural resource conservation. The NWCC houses data acquired from NWS' COOP network, as well as data acquired with the Agency's own data collection networks, Snow Telemetry (SNOTEL) and Soil Climate Analysis Network (SCAN). Both of these networks are comprised of automated remote sites that transmit data as frequently as hourly to the Centralized Computer Facility in Portland, OR. This facility and the associated data collection and dissemination are a 24/7 operation, with minimal downtime allowed. The center supports data management, data analysis and product generation. The Center is also responsible for developing and acquiring other water resource related technology for use by NRCS and its Partners and cooperators.

5.5. FSA – Kansas City, MO

The FSA Kansas City Management Office (KCMO) provides the IT development, operations and maintenance support for FSA field offices. FSA KCMO serves as FSA's primary web farm and is currently developing a small number of geospatial applications including the Service Center Information Management System (SCIMS) and Commodity Reporting Tool.

6. USDA Information Technology Organizations

As the role of each data center is defined as part of the Geospatial Data Warehouse community, it is necessary to interact with USDA's IT organizations and agencies that lie outside of the geospatial community. This interaction will ensure that the geospatial data centers will implement USDA architectures and adopt department recommendations that will ensure the continued growth and health of the geospatial data warehouse by gaining the early support from the broader USDA IT community. The current work of two of USDA's IT groups is presented here, the Common Computing Environment (CCE) and the Electronic Access Initiative (EAI). A more detailed summary of the architecture recommendations and impacts to the geospatial data warehousing initiative are presented in **Appendix A**.

6.1. Common Computing Environment (CCE)

The Common Computing Environment (CCE) Information Technology Architecture (ITA) Version 5, September 2001, recommends maintaining the data ownership model currently employed at the national level and at local level Service Centers. This distributed data ownership model results in the need to replicate nationally owned data locally and locally owned data replicated nationally. The ITA continues to describe the intermediate (over the next three years) and long term (beginning in four or more years) plans for expansion and development of distributed services architecture. These plans are described in more detail in **Appendix A**.

6.2. Electronic Access Initiative (EAI)

The Electronic Access Initiative (EAI) is responsible for implementation of USDA's Web Farms. Web Farms play an important role as hosts for the applications that provide data access and dissemination services within the geospatial data warehouse. Therefore, it is important to consider the current architecture of the Web Farms and how this configuration will impact the implementation of the geospatial data warehouse. Current EAI architectures and the impacts they have on the geospatial data warehouse are presented in **Appendix A**.

7. External Impacts and Influencing Technologies

As USDA implements the geospatial data architecture plan, external agencies and organizations are moving forward in the development of standards and specifications that will impact USDA Geospatial Data Warehouse in the future. These influencing technologies are important to USDA's long-term role as a provider of certified source geospatial data, implementation of E-Gov initiatives, and maintaining a presence as an interoperable member of the geospatial community.

7.1. NSDI

Goal 2 of the National Spatial Data Infrastructure (NSDI) Strategy¹² is to “*Develop common solutions for discovery, access, and use of geospatial data in response to the needs of diverse communities.*” The four objectives of that goal are to:

1. Continue to develop a seamless National Geospatial Data Clearinghouse.
2. Support the evolution of common means to describe geospatial data sets.
3. Support the development of tools that allow for easy exchange of applications, information, and results.
4. Research, develop, and implement architectures and technologies that enable data sharing.

Goals 1 and 2 are well underway within USDA especially at NRCS/NCGC. As one of the initial NSDI search nodes, NCGC has demonstrated leadership in the collection, maintenance, and distribution of metadata to support the NSDI. Goals 3 and 4, however, represent an emerging national effort that USDA could contribute to significantly. Work currently underway among national and international standards bodies as well as industry will have a major impact on USDA’s ability to contribute to and take advantage of the interoperability tools, architectures and technologies described in the goals above.

7.2. OpenGIS

The primary mission of the OGC is to address the lack of interoperability among systems that process geo-referenced data, specifically addressing disparate and often distributed geo-based computing systems. The OGC does this by coordinating the development of industry-wide interface specifications that will lead to standards based commercial software.

USDA plays a critical role in the OGC as sponsor and test bed participant. Participation at this level ensures that USDA requirements influence the OGC specifications and, in turn, drive the technology development within architectural parameters of USDA and their partner organizations. More specifically, USDA can now choose to implement standards-based commercial-off-the-shelf-software (SCOTS) with the prospect of interoperating with their partners across several interfaces.

Figure 7-1 illustrates this concept at a high level. Of current interest to USDA are the SCOTS products developed or currently under development by ESRI, USDA’s enterprise GIS provider.

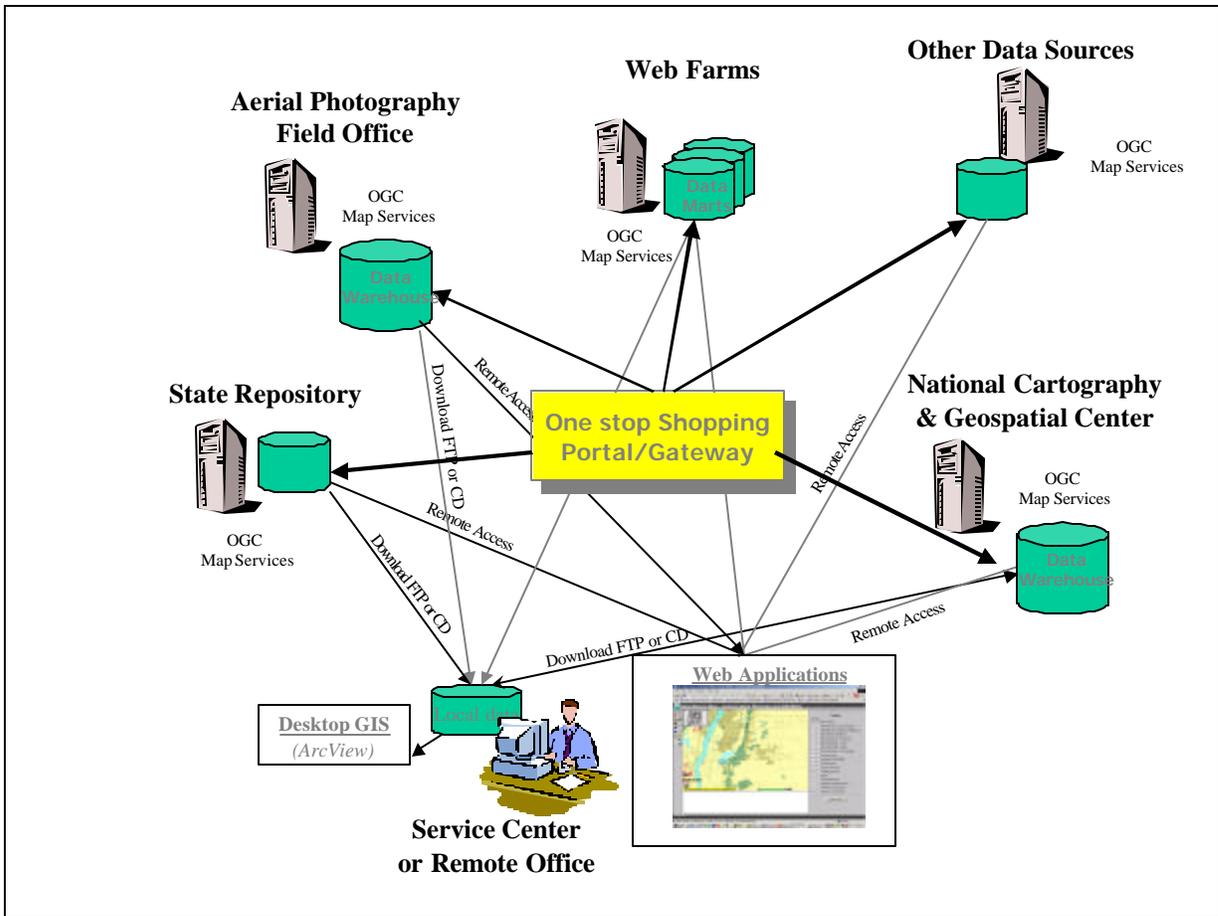


Figure 7-1 OpenGIS technology to support the USDA Data Warehouse

7.3. Geospatial One-Stop

The OMB Geospatial One-Stop was mentioned as a business case driver earlier in this document. If implemented, this approach could have a profound impact on the USDA Geospatial Data Warehouse architecture from both a consumer and producer perspective. As a producer of information, USDA does not have direct stewardship responsibilities for any of the framework layers identified by One-Stop; however, USDA is a significant contributor in the funding of both the elevation and ortho framework layers. It is also conceivable that USDA soils information could be added to the list of framework layers. From a data consumer perspective, USDA’s extensive presence in local communities and need for geospatial information will certain drive much of the requirements for local, state, and national One-Stop participation.

7.4. Environmental Systems Research Institute (ESRI)

In 2001, ESRI was selected as the GIS vendor of choice for USDA and was awarded an enterprise agreement to supply geospatial technology, services and training for all of USDA. This agreement will likely have a profound effect on the Geospatial Data Warehouse architecture of USDA as well as an influencing effect on the technology

direction of ESRI. Within recent years ESRI has gone through a technology evolution that focused on a more integrated and standards based architecture. ESRI appears to be concentrating efforts in the Microsoft .NET technology arena to support their distributed geospatial service offerings. As the ESRI architecture solidifies, USDA is likely to adopt a similar architecture.

7.5. Other Governmental and Commercial geospatial data warehouses

There are many reference data warehousing implementations that could be reviewed as best practices, however, the number that integrate spatial information as a critical part of the information model are few. The Environmental Protection Agency (EPA) is one of those few.

EPA EnviroFacts data warehouse is one successful federal implementation of a spatially enabled data warehouse. The EnviroFacts data warehouse has been online since 1995. From the EPA's six national mainframe systems, the warehouse application extracts information on approximately one million sites handling or discharging potentially harmful substances and pulls regulatory, spatial and demographic data into a warehouse built on software from Oracle and ESRI. Data from the 40 GB warehouse is accessible to anyone with a Web browser as illustrated in Figure 7-2. Users include EPA staff members, emergency management teams, business and industry, community advocates and real estate brokers.



Figure 7-2 EnviroMapper Interface

Pat Garvey, director of the Environmental Protection Agency's EnviroFacts data warehouse team, offered the following lessons learned:¹³

- "Don't think too big, and don't start off too grandiose. Keep expectations lowered."
- "The project manager has to have a vision that's direct and flexible. It's difficult to keep expectations lowered and keep people engaged. So you have to keep on making

incremental successes very quickly throughout the project life. The mini-success stories are just crucial.”

-- “Always plan for expansion and enhancement—expansion because you want to keep driving more data into the warehouse and enhancement because you want to be able to access easily and completely the data within the warehouse,”

-- “If you’re using different computer-aided software engineering tools, different documentation tools and different database modeling tools, then you’re in a situation where seamlessness isn’t really there. Staying within the vendor family diminishes a lot of the chaos of matching up tools and other vendors.”

8. Technology Review

In order to fulfill the business and technical requirements of the geospatial data warehouse architecture presented in Section 4, certain technologies will need to be assessed, procured, and implemented as needed. In addition to ensuring that data are available, data must be locatable, accessible and secure. The technologies that will be implemented to meet these requirements are presented here. In many cases, several competing technologies are presented and the benefits of each are compared.

8.1. Availability

As USDA invests in richer and more accurate spatial data, the need for high data availability solutions becomes more profound. There are many facets and technology solutions to achieving high availability, each with associated cost and risk mitigation. These include technologies such as redundancy, disaster recovery and fail-over. Most if not all of these technologies apply to all USDA business data, not just geospatial and should be adequately addressed by the OCIO and their technical offices (i.e. EAI, CCE, Web Farms). However, only a high-level discussion is appropriate for this study. Of particular concern are the resources (human, technical) required to achieve high geospatial data availability for business applications. When determining the cost-benefit of such solutions, the following should be considered.

8.1.1. Redundancy

Consider the number of components that must be operational in order to deliver a typical on-line USDA GIS application. A complex configuration including a database, storage device, application server, web server, network, routers, switches, and power supply must all work together in order to deliver the application; not to mention the staff resources required to maintain and operate the equipment. If any of those components should fail due to overloaded use, equipment failure, or catastrophic event then the business will not get done.

8.1.1.1. Fail over

One approach is to eliminate single points of failure in the system. A holistic approach to this is probably the best. This includes

- **multiple web servers** – on-line GIS applications can be running on multiple machines simultaneously. This will ensure that one or more are available in the event of a failure and can provide load balancing during periods of heavy use. F5's BIG-IP is an example of a product that provides this service. Using BIG-IP, a server cluster appears to its clients as a single server but is, in fact, a group of servers acting as one.
- **redundant storage devices** - both APFO and NCGC currently employ a level of redundant storage. APFO's solution relies on a near on-line tape storage device that serves up files to disk storage on an as-needed basis. Both APFO and NCGC are currently relying on Redundant Array of Independent Disks (RAID) storage to recover data from inadvertent disk failure. Each center could become a backup storage location for the other to ensure greater level of redundancy, however, replicating data between data centers has significant additional costs associated with it.
- **clustered and “host swappable” CPUs** – APFO and to a lesser extent NCGC uses the Sun platform for their primary storage. Many of their newer servers are capable of swapping failed CPUs without bring down the machine. This allows for continued, but possibly reduced performance during server maintenance operations.
- **redundant network paths and equipment** – Currently all USDA web map services are running at NRCS ITC and NCGC. At NCGC, these services are dependant on a single T1 network connection and router system at one location in the building. Any catastrophic event to the network line or would result in loss of business. Multiple lines and OCIO Stack equipment could be installed in different parts of the facility to prevent this from happening. However, there is significant cost in implementing redundant network equipment and should be weighed against the cost-benefit of implementing the same configuration at an additional facility.
- **multiple data center facilities** - The identification for alternative or back-up data centers hold much potential in reducing the risk to business continuance. However, this can be an expensive proposition whether the alternatives are other USDA locations or outsourced data centers that share IT resources with other data center customers. Distance, connectivity and staffing issues should be considered prior to determining where such a center should be located and whether or not an alternative data center is cost effective.

8.1.1.2. Data Replication

In general, data replication can be categorized¹⁴ as:

Transaction-Aware Replication: Transaction-aware replication offers transaction-level replication, typically by electronically transmitting database or file changes (i.e., through

logs) to the secondary site and applying those changes to a replica image. The primary advantage of this approach is that the replication method understands units of work (i.e., transactions) and has a greater potential for data integrity (via transaction roll-forward/back), although data integrity is not guaranteed. The downside of this approach is performance, especially when replicating large databases such as spatial databases.

Mirroring or Shadowing: Shadowing maintains a replica of databases and/or file systems, typically by continuously capturing changes and applying them at the recovery site. Shadowing is an asynchronous process, thus requiring less network bandwidth than synchronous mirroring. Recovery time objectives (RTOs) are significantly reduced (generally between one and eight hours, depending on the lag time for applying logs), while recovery point objectives (RPOs) are as up-to-date as the last receipt and application of the logs.

Mirroring maintains a replica of databases and/or file systems by applying changes at the secondary site in lock step with or synchronous to changes at the primary site. As a result, RTOs can be reduced to between 20 minutes to several hours, while RPOs are reduced only to the loss of uncommitted work. Because it is synchronous, mirroring requires significantly greater network bandwidth than shadowing. Too little bandwidth and/or high latencies will degrade the performance of the production system. Some of the factors driving these various solutions are:

- Distance from source to target data store
- Commonality of platform between source and target data store
- Type and volume of data to be mirrored
- Frequency of mirroring

Data replication has a variety of applications in the data-warehousing environment. The following lists some common requirements for data replication:

- May respond to the need to have a secure backup of the data assets in case the primary site fails.
- Allows for synchronization of multi-node access to the same data for load balancing purposes.
- Allows for the rapid integration of data from production sites to the data warehouse or from the data warehouse to the data marts used for analysis.

Moreover, there are quite a number of technical approaches to data replication depending on the need. It is unlikely that one “silver bullet” approach to data replication will satisfy all business and technical requirements of the system. In fact, it may be more prudent to recognize the pros and cons of each method and use them accordingly. The following list illustrates the variety of approaches to implementing data replication and the number of vendors that provide those solutions. A table included in Appendix B reprinted from a Gartner Group Research Note¹⁴, presents data replication options with some differentiating features.

USDA must implement a replication solution that satisfies the need to copy replicates of potentially large volumes of spatial data from a primary to a secondary data warehouse as well as from the warehouse to the data marts wherever they may be. Table 8-1 presents some technologies and methods to consider. This list is not exhaustive, but represents the technology that is most commonly deployed at each of the potential data warehousing sites.

Table 8-1 Alternative Replication Technologies

Replication Type	Company/Product	Characteristics
Transaction Aware	Oracle 8i	<ul style="list-style-type: none"> • Basic replication is implemented using standard CREATE SNAPSHOT or CREATE MATERIALIZED VIEW statements. It can only replicate data (not procedures, indexes, etc), replication is always one-way, and snapshot copies are read only. • Advanced replication supports various configurations of updateable-snapshot, multi-master and update anywhere replication. It is more difficult to configure, but allows data and other database objects like indexes and procedures to be replicated. • Cannot replicate LONG and LONG RAW data types (Use LOBs instead)
	ESRI ArcSDE	<ul style="list-style-type: none"> • ESRI supports this on Oracle 8.1 and SQLServer2000. • DBMS replication should be supported by Informix and DB2 in the current or next release. • Each DBMS implementation has a unique implementation of replication and MAY only work in a homogeneous DBMS environment. • ESRI conducted a test-bed on this method and found it to be insufficient for most use cases. • ESRI recommended an alternative called “Procedural Based” which involves the export and import of data from one ArcSDE database to another.
Mirroring or Shadowing	Sun StorEdge Instant Image	<ul style="list-style-type: none"> • Takes instant, point-in-time snapshots of online data and makes them available to another application running on the primary or a secondary host. • Keeps track of the differences between the master volumes and your shadow volumes. Only the blocks that are changed need to be

		updated to resynchronize master and shadow volumes.
	Sun Network Data Replicator (SNDR)	<ul style="list-style-type: none"> • Employs an open architecture, may be used to replicate third-party storage, and can be combined with Sun StorEdge Instant Image software for additional business continuance services. • Allows replication of data in both synchronous and asynchronous modes. • Can replicate any third-party storage. • primary volumes can be protected by any RAID level desired (RAID 1, RAID 5, etc.). The protection level of the secondary volumes does not have to match that of the primary volumes.
	EMC Symmetrix Remote Data Facility (SRDF)	<ul style="list-style-type: none"> • Provides simple, platform-independent, server-less data replication across business and application boundaries. • Enables simultaneous information sharing across multiple locations and geographies. • Provides the ability to distribute production copies of data for application testing or development. • Offers content distribution without impact on server or application cycles regardless of the volume of data involved.

8.2. Web Services

Web services are a new breed of Web application. They are self-contained, self-describing, modular applications that can be published, located, and invoked across the Web. Web services perform functions, which can be anything from simple requests to complicated business processes. Once a Web service is deployed, other applications (and other Web services) can discover and invoke the deployed service. As mentioned, both ESRI and OpenGIS are beginning to develop interfaces and software based on web service technology. In addition, many of the popular database vendors (i.e. IBM and Microsoft) are promoting the use of web service technology in the USDA Service Center Agency IT centers. Web service technology enhances the Geospatial Data Warehouse Architecture by providing common interface framework between the various components of the warehouse. As web service standards such as XML, Web Service Description Language (WSDL), Universal Description, Discovery, and Integration (UDDI) and Simple Object Access Protocol (SOAP) mature, it is likely that those technologies and their successors will become commonplace in the architecture.

9. Implementation Plan – A Phased Approach

9.1. Near-term Implementation Phase

Much of the work currently underway within the agencies data warehousing efforts, the Resource Data Gateway and the Data Centers lays a solid foundation for future data warehousing activities. To move forward from this point, the USDA must begin to solidify the common architecture that will support this effort on an enterprise scale. However, many questions remain to be answered. One way forward would be to conduct a proof-of-concept pilot to:

- Provide additional data to support cost-benefit analysis and implementation of data warehousing technologies to support USDA business needs
- Demonstrate the feasibility of candidate technologies
- Refine the specifications acquiring data warehousing technology

Table 9-1 illustrates the crosswalk of business drivers identified in Section 4 pilot objectives that could be addressed within a short series of proof-of-concept pilots. Not all pilots are necessarily required immediately. They should also be prioritized in order of criticality to support the business.

Table 9-1 Business Driver to Technical Requirement to Pilot Crosswalk

Business Requirement	Technical Requirement	Pilot Objective
Implement privacy and security requirements by data set	Restrict Access to information using industry standard network security techniques	Pilot various technologies to determine the appropriate level of protection for geospatial information including: ArcIMS Access Control Lists, SSL, and VPN/PKI
Define Metrics for Administration	Implement industry standard network monitoring solution	Pilot an industry standard solution for network monitoring
Provide Feedback Mechanism for Users	Users may either subscribe to data notification service that indicates when data in their area of interest has been updated and/or a data content catalog not allows users to check the currency of data through a common interface.	Pilot data notification (push/pull) system for when elements of the data warehouse are updated. Determine the behavior characteristics of both
Determine access	Implement standard	None

requirements for clients (audience, speed for unique client data)	documentation and training for geospatial data warehouse	
Support delivery in multiple formats as requested by user (within limits, structure remains unchanged)	Warehouse transformation processes must support the following transformations: Binary transformation Projection transformation Generalization Aggregation Interpolation Interpretation	Pilot the top priority transformations required by the business
Ability to incorporate full support of metadata	Implement a metadata management and publication solution	Pilot the use of various solutions including: Compusult MetaManager on the ESRI ArcIMS Metadata Server
Framework for providing common services (objects, COM, Net tech, etc.)	Develop technical framework based on industry standards that allows consistent deployment of business applications within the framework	Outline a technical architecture framework that accommodates the heterogeneity of the current multi-agency environment
Ability to provide high availability of data with reliability and integrity of data intact	Implement strategies for load balancing, fail-over, and disaster recovery	Pilot using industry standard technology
Replication should be automated to minimize human intervention	Implement data replication at various points in the architecture to support high availability needs	Pilot a variety of data replication solutions including: ArcSDE Data Replication RDBMS Replication in both heterogeneous and homogeneous DBMS environments (Oracle, SQLServer, Informix, DB2) Device level replication from storage vendors such as Sun and/or EMC
Develop policy and procedures for state level data set replication at the state data marts.	Involve State agencies in the deployment of data marts throughout the architecture	Pilot State level data mart

Facilitate acquisition, integration, and distribution of warehouse data	Implement a variety of COTS and potentially customized ETL tools for populating the data warehouse from operations data stores and for populating data marts from data warehouses and enterprise level data bases	Pilot various ArcSDE, DBMS or data warehousing software for migration of data throughout the system
Support versioning configuration for online and offline migration	Implement data warehouse structure that permits the storage and retrieval of historic geospatial information to support a variety of time-sensitive analysis	Pilot the ArcSDE Geodatabase vs. typical RDBMS record level vs. file level versioning to support various temporal analysis
Support OGC standards, both currently adopted and future considerations. This includes the ability to integrate multiple geodatabases in multiple data stores	Implement OGC standards based components that are applicable to the business needs of the users	Pilot the implementation of WMS Services, Catalog Services, and clients to support these services as well as OpenGIS access from other partner agencies
System should be user friendly KISS, but meet the needs of more knowledgeable customers and support custom application development.	N/A	N/A
Integrate data to SC standards	Must support the standards defined for the management of geospatial data at the service center	Pilot the automatic transformation of data from the warehouse to the standard configuration of data on the local machine
Address user requirement to deliver integrated data in specific geographical regions	Support the creation of data marts for any USDA business application in any region supported by USDA	Pilot the creation of subject and State specific data marts using various technologies such as replication
Identify and access current official data for a specific use	Support the near real-time flow of information from operational data stores through the warehouse to the data marts and end users	Pilot the flow of integrated information from the soils warehouse perspective

The business drivers presented in Table 9-1 have been migrated into a preliminary pilot project plan presented in Figure 9-1. The goal of the pilot is to implement geospatial data warehousing within four months, building towards the long-term development and O&M plan.

9.2. Geospatial Data Warehouse Support Service Roles and Responsibilities

During the examination of the composition of the geospatial data warehouse components it is important not to overlook those human resources that are ultimately responsible for the successful implementation and maintenance of the warehouse. It is these resources that are responsible, not only for the day-to-day technical systems administration of the data centers and web farms, but the people responsible for defining the standards, requirements and applications that will be employed on these systems and in these centers.

Table 9-2 Geospatial Data Warehouse Roles and Responsibilities

Role	Responsibility
Project Manager	<ul style="list-style-type: none"> ✓ General oversight and responsibility for the delivery of system services in a production environment. ✓ Includes the acquisition and managing of resources, coordination with the system's Executive Sponsor, monitoring of production statistics, and general system management responsibilities.
Warehouse Architect	<ul style="list-style-type: none"> ✓ Primary administrative responsibility for the actual physical design of the warehouse environment ✓ Participates in the modeling activities as the representative of the physical implementation of the model's entities ✓ Oversees the creation of the database tables and the maintenance of the warehouse's physical environment and monitors the changes made to the environment by more junior database analysts ✓ Maintains the vision of the physical view of the warehouse.
Data Modeler	<ul style="list-style-type: none"> ✓ Builds the models, and then validates them by mapping data from the operational models to the model of the warehouse.
Senior Database Admin.	<ul style="list-style-type: none"> ✓ Minimizing redundant operational data and structuring data to serve the organization's overall purpose in capturing that data. ✓ Develop and publish naming standards, data element mapping procedures and guidelines for database creation. ✓ Tuning and maintaining the operations of the DBMS platform.

Junior Database Admin.	<ul style="list-style-type: none"> ✓ Provides basic database support including managing accounts, loading tables, backup, and recovery.
Senior Programmer	<ul style="list-style-type: none"> ✓ Extraction of specific data elements from various systems, databases and files into a relational database ✓ Changing the formats of certain elements to conform to a standard, making all iterations of an element have the same data type and size or altering the algorithm that summarizes certain elements ✓ Develops analytical and computational techniques and methodology for problem solutions. ✓ Performs enterprise-wide strategic systems planning, business information planning, business and analysis. Performs process and data modeling in support of the planning and analysis efforts using both manual and automated tools, such as Integrated Computer-Aided Software Engineering (I-CASE) tools. ✓ Applies reverse engineering and re-engineering disciplines to develop migration strategic and planning documents.
Junior Programmer	<ul style="list-style-type: none"> ✓ Analyzes application software and design specifications for information process activities. ✓ Develops block diagrams and logic flow charts. ✓ Translates detailed design into application software. ✓ Tests, debugs, and refines the application software to produce the required product. ✓ Prepares required documentation, including both program-level and user-level documentation. May serve as Tier 1 support to users and involved with trouble-shooting system operations.
Telecom/Security Spec.	<ul style="list-style-type: none"> ✓ Implements and manages Internet and Intranet security procedures, monitoring system security breaches and notifies authorities of unauthorized access. ✓ Monitor the implementation of security update/patches as needed. Additionally, they ensure the rigorous application of information security/information assurance policies, principles, and practices.
Web Admin. (3 on-site)	<ul style="list-style-type: none"> ✓ Responsible For Implementation And Management Of Internet Web Services. ✓ Maintaining Web Servers, Web Software, Telecommunications Connectivity, Monitoring Web Site Functionality, And Integrity, Troubleshooting And Resolving Problems, Reviewing, Testing, And Integrating Web Pages, Collecting And Analyzing Web Site Statistics.
Training Spec.	<ul style="list-style-type: none"> ✓ Develop the materials they will use in these sessions. ✓ Areas for education should include the business case

	for a data warehouse, the differences between developing operational systems and warehouses, the need for strong data administration (also called information resource management) in building and maintaining a data warehouse.
Metadata Spec.	<ul style="list-style-type: none"> ✓ Single point of access and control for an enterprise's meta data ✓ Meta data indicates where the data comes from, how it should be translated or transformed, its form and functions. Serves as the liaison between the technical and user communities for the operational and warehouse meta data ✓ Manages the integration of the logical models of the operational and warehouse systems ✓ Participates on the standards development team.
Technical Support (4 on-site)	<ul style="list-style-type: none"> ✓ This function includes hardware support. ✓ Manages daily backup processes, technology refreshes, enforcement of system security, coordinating with communications and other service providers, and general system management and maintenance. ✓ Planning and scheduling the installation of new or modified hardware/software, allocating systems resources, managing accounts and resolving hardware/software interface and interoperability problems. ✓ Producing CDs ✓ Loading new data sets
Vendor Consulting	<ul style="list-style-type: none"> ✓ Provides vendor specific insight and guidance on the implementation and integration of COTS products. ✓ Acts as liaison to vendor development staff for technical and engineering change request support.

9.3. Geospatial Warehouse Pilot Schedule and Cost

It has long been known that the data warehouse environment, with its many components, is best developed in an iterative manner.¹⁵ Figure 9-1 identifies the high-level work breakdown structure and schedule for a four month Geospatial Data Warehouse Pilot effort. The schedule incorporates government, support services, and vendor tasks aimed at providing proof-of-concept and risk mitigation for the warehouse project.

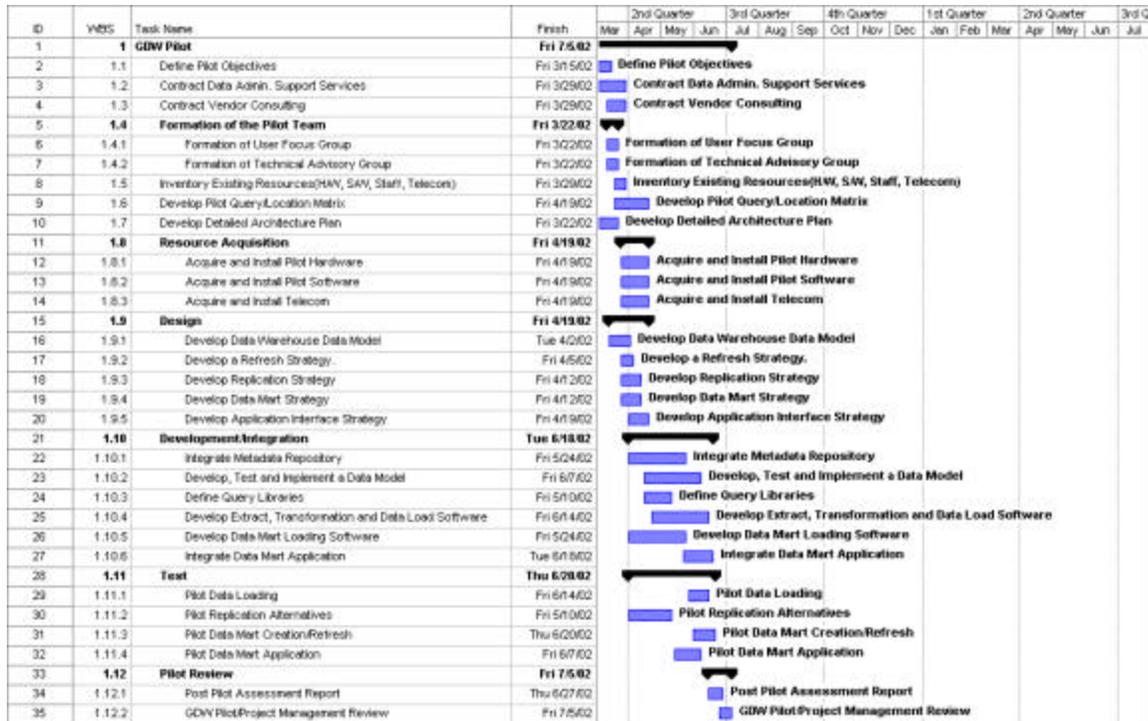


Figure 9-1 Pilot Geospatial Data Warehouse Implementation Schedule

Table 9-3, Table 9-4, and Table 9-5 are a summary of the projected support service costs for the Geospatial Data Warehouse Pilot and Phase I & II iterations. Labor categories bolded in red indicate those roles and responsibilities that carry over into future operations and maintenance phases. Phase II is will begin the operations and maintenance phase of the warehouse activity. Iterative development phases will continue beyond phase II as required by the business community.

Table 9-3 Pilot Geospatial Data Warehouse Support Service Labor Costs

	Rate	Geospatial Data Warehouse Pilot				Total Hours	Total Cost
		Mar-02	Apr-02	May-02	Jun-02		
Project Manager	\$85	80	80	80	80	320	
		\$ 6,800	\$ 6,800	\$ 6,800	\$ 6,800		\$ 27,200
Warehouse Architect	\$100	160	120	80	80	440	
		\$16,000	\$12,000	\$ 8,000	\$ 8,000		\$ 44,000
Data Modeler	\$60	160	160	160	160	640	
		\$ 9,600	\$ 9,600	\$ 9,600	\$ 9,600		\$ 38,400
Senior Database Admin.	\$65	0	0	0	0	0	40
		\$ -	\$ -	\$ -	\$ -		\$ -
Junior Database Admin.	\$50	0	0	0	0	0	0
		\$ -	\$ -	\$ -	\$ -		\$ -
Senior Programmer	\$75	80	160	160	160	560	20
		\$ 6,000	\$12,000	\$12,000	\$12,000		\$ 42,000
Junior Programmer	\$50	0	0	0	0	0	20
		\$ -	\$ -	\$ -	\$ -		\$ -
Telecom/Security Spec.	\$60	0	120	80	20	220	40
		\$ -	\$ 7,200	\$ 4,800	\$ 1,200		\$ 13,200

Web Admin. (3)	\$55	160	160	160	160	640	
		\$ 8,800	\$ 8,800	\$ 8,800	\$ 8,800		\$ 35,200
Training Spec.	\$50	0	0	0	0	0	0
		\$ -	\$ -	\$ -	\$ -		\$ -
Metadata Spec.	\$65	0	0	0	0	0	80
		\$ -	\$ -	\$ -	\$ -		\$ -
Technical Support (4)	\$45	0	0	0	0	0	
		\$ -	\$ -	\$ -	\$ -		\$ -
Vendor Consulting	\$160	80	80	0	0	160	
		\$12,800	\$12,800	\$ -	\$ -		\$ 25,600
						2980	\$225,800

Table 9-4 Phase I Geospatial Data Warehouse Support Service Labor Costs

	Rate	Geospatial Data Warehouse Phase I					Total Hours	Total Cost
		Jul-02	Aug-02	Sep-02	Oct-02	Nov-02		
Project Manager	\$85	160	160	160	160	160	800	
		\$13,600	\$13,600	\$13,600	\$13,600	\$13,600		\$ 68,000
Warehouse Architect	\$100	160	160	120	80	40	560	
		\$16,000	\$16,000	\$12,000	\$ 8,000	\$ 4,000		\$ 56,000
Data Modeler	\$60	160	160	120	80	40	560	
		\$ 9,600	\$ 9,600	\$ 7,200	\$ 4,800	\$ 2,400		\$ 33,600
Senior Database Admin.	\$65	120	120	80	80	40	440	
		\$ 7,800	\$ 7,800	\$ 5,200	\$ 5,200	\$ 2,600		\$ 28,600
Junior Database Admin.	\$50	0	160	160	160	160	640	
		\$ -	\$ 8,000	\$ 8,000	\$ 8,000	\$ 8,000		\$ 32,000
Senior Programmer	\$75	120	160	160	160	160	760	
		\$ 9,000	\$12,000	\$12,000	\$12,000	\$12,000		\$ 57,000
Junior Programmer	\$50	120	160	160	160	160	760	
		\$ 6,000	\$ 8,000	\$ 8,000	\$ 8,000	\$ 8,000		\$ 38,000
Telecom/Security Spec.	\$60	40	40	20	20	20	140	
		\$ 2,400	\$ 2,400	\$ 1,200	\$ 1,200	\$ 1,200		\$ 8,400
Web Admin. (3)	\$55	160	160	320	320	480	1440	
		\$ 8,800	\$ 8,800	\$17,600	\$17,600	\$26,400		\$ 79,200
Training Spec.	\$50	20	20	20	120	160	340	
		\$ 1,000	\$ 1,000	\$ 1,000	\$ 6,000	\$ 8,000		\$ 17,000
Metadata Spec.	\$65	160	160	160	120	80	680	
		\$10,400	\$10,400	\$10,400	\$ 7,800	\$ 5,200		\$ 44,200
Technical Support (4)	\$45	160	160	320	320	320	1280	
		\$ 7,200	\$ 7,200	\$14,400	\$14,400	\$14,400		\$ 57,600
Vendor Consulting	\$160	80	40	40	20	0	180	
		\$12,800	\$ 6,400	\$ 6,400	\$ 3,200	\$ -		\$ 28,800
							8400	\$462,000

Table 9-5 Phase II Geospatial Data Warehouse Support Service Labor Costs

		Geospatial Data Warehouse Phase II								
	Rate	Dec-02	Jan-03	Feb-03	Mar-03	Apr-03	May-03	Jun-03	Total Hours	Total Cost
Project Manager	\$85	160	160	160	160	160	160	160	1120	
		\$ 13,600	\$13,600	\$13,600	\$13,600	\$13,600	\$13,600	\$13,600		\$ 95,200
Warehouse Architect	\$100	40	40	20	0	0	0	0	100	
		\$ 4,000	\$ 4,000	\$ 2,000	\$ -	\$ -	\$ -	\$ -		\$ 10,000
Data Modeler	\$60	120	120	80	80	40	0	0	440	
		\$ 7,200	\$ 7,200	\$ 4,800	\$ 4,800	\$ 2,400	\$ -	\$ -		\$ 26,400
Senior Database Admin.	\$65	120	80	80	40	40	40	40	440	
		\$ 7,800	\$ 5,200	\$ 5,200	\$ 2,600	\$ 2,600	\$ 2,600	\$ 2,600		\$ 28,600
Junior Database Admin.	\$50	160	160	160	160	160	160	160	1120	
		\$ 8,000	\$ 8,000	\$ 8,000	\$ 8,000	\$ 8,000	\$ 8,000	\$ 8,000		\$ 56,000
Senior Programmer	\$75	120	80	40	40	40	40	40	400	
		\$ 9,000	\$ 6,000	\$ 3,000	\$ 3,000	\$ 3,000	\$ 3,000	\$ 3,000		\$ 30,000
Junior Programmer	\$50	160	160	160	160	160	160	160	1120	
		\$ 8,000	\$ 8,000	\$ 8,000	\$ 8,000	\$ 8,000	\$ 8,000	\$ 8,000		\$ 56,000
Telecom/Security Spec.	\$60	20	20	0	0	0	0		40	
		\$ 1,200	\$ 1,200	\$ -	\$ -	\$ -	\$ -	\$ -		\$ 2,400
Web Admin. (3)	\$55	480	480	480	480	480	480	480	3360	
		\$ 26,400	\$26,400	\$26,400	\$26,400	\$26,400	\$26,400	\$26,400		\$184,800
Training Spec.	\$50	20	0	0	20	20	80	160	300	
		\$ 1,000	\$ -	\$ -	\$ 1,000	\$ 1,000	\$ 4,000	\$ 8,000		\$ 15,000
Metadata Spec.	\$65	120	120	80	80	80	80	80	640	
		\$ 7,800	\$ 7,800	\$ 5,200	\$ 5,200	\$ 5,200	\$ 5,200	\$ 5,200		\$ 41,600
Technical Support (4)	\$45	640	640	640	640	640	640	640	4480	
		\$ 28,800	\$28,800	\$28,800	\$28,800	\$28,800	\$28,800	\$28,800		\$201,600
Vendor Consulting	\$160	20	20	0	0	0	0		0	
		\$ 3,200	\$ 3,200	\$ -	\$ -	\$ -	\$ -	\$ -		\$ 6,400
									13560	\$546,000

Table 9-6 summarizes the entire budget requirement from Pilot to Phase II of the Geospatial Data Warehousing effort. With the exception of additional storage and application server capacity much of the piloting can be completed using the existing data center and web farm infrastructure. However, a significant amount of infrastructure investment will be required to support the initial phases of the effort.

Table 9-6 Geospatial Data Warehouse (Pilot - Phase II) Budget

INFRASTRUCTURE ITEM	Pilot	Phase I	Phase II
Storage (Online)		\$294,000	\$294,000
Storage (Near-Online)	\$100,000	\$550,000	\$550,000
Servers (data, web, applications.)	\$100,000	\$2,000,000	\$1,350,000
Other Hw/Sw		\$250,000	\$80,000
Data Mgt Software		\$1,000,000	\$900,000
ETL/OLAP Software		\$550,000	\$550,000
Telecommunications \$/yr		\$0	\$250,000
Telecommunication Security		\$210,000	\$0
Physical Security		\$0	
Support Services - \$/yr	\$225,800	\$462,000	\$87,800
Implementation of Servers, communications, and replication			
O&M (hardware, software, staffing,			\$666,200
TOTAL	\$425,800	\$5,556,000	\$4,728,000

9.4. Long-term Implementation Phases

Update the Architecture

As each pilot activity is conducted, the results must be assessed as to their satisfaction of the business need. In some cases, the pilot will lead directly into national implementation whereas other pilots might instigate the need for further requirements refinement or solution trade-offs. In either case, the results of the pilots must be used to refine the overall common architecture of the data warehouse.

Drive the Architecture

To date, much of the technical architecture of the CCE is driven by requirements derived from a primarily stand-alone, distributed architecture. The Geospatial Data Management Team must get ahead of the deployment decisions and help define the future procurements rather than retrofit within the existing configuration.

Institutionalize the Activity

The efforts of the Geospatial Data Management Team have been largely confined within the walls of ITC and the Geospatial Data Production Centers, whereas other organizations (CCE, EAI, Web Farms) lead the way in setting the course of USDA IT. The Geospatial Data Warehousing activity must become more a part of the overall integrated USDA IT institution in order make a significant contribution to USDA business goals.

Cost Realism

The previous study put forth a high level cost-benefit designed to assess the trade-off between centralized and distributed geospatial data warehouses. By design, the costs were decremented by the resources (technical & human) already in place at the various

locations (APFO, NCGC, Web Farms, Data Gateway). That presumes that those resources would be allocated to the data warehouse task regardless of the outcome. This approach, although accurate in some cases, tends to underestimate the cost of implementation and ignore the need to budget for a procure resources specific to the task. It is important to distinguish a cost benefit analysis (CBA), which is an investment decision tool, from a budget, which is a resource allocation tool. The CBA incorporates investment analysis methods to identify financially superior alternatives.

10. Summary

In summary, there is a strong business case and legislative mandate for moving forward on implementation. If USDA is expected to delivery integrated programs to their customers and partners, then we must recognize first, that much of that integration is based on geography and second, that geospatial data warehousing is the technology that enables ubiquitous geospatial and program data integration across the enterprise.

Appendix A - Current Data Centers Architecture

APFO Architecture

The Aerial Photography Field Office is the primary acquisition source of base imagery for USDA agencies, the Farm Service Agency (FSA), National Resources Conservation Service (NRCS), and the Forest Service (FS). Geospatial data and imagery is managed by the APFO for integration and use by USDA Field Service Centers in GIS. Additional image acquisition and reproduction requirements include program compliance, resource inventories, resource planning, and disaster assessment. The APFO continues to provide leadership for long-range operating policies, program initiatives, data standards, and technological advancements in the fields of photogrammetry, remote sensing, aerial photography, digital imaging, and geospatial data.

Hardware

Production hardware at the APFO is dominated by Sun servers and workstations and is driven by application software to support digital mosaic production. There are a variety of imagery processing tasks supported by this environment, from cataloging of imagery to reproduction services. Digital imagery production activity is supported by a 2.8 TB disc array and a 195 TB capacity near on-line tape storage system. The APFO utilizes NT workstations for digital imagery and CLU inspection processes.

Software

The APFO has traditionally used Oracle for RDBMS support of 8 management/production systems and recently upgraded to Oracle 8i. LH Systems' SocetSet lines of products are used for digital imagery rectification, mosaicking and image enhancement. Multi-Resolution Seamless Image Database (MrSID) is used for image compression. The APFO recently received the full suite of ESRI products as a result of the enterprise agreement reached with the USDA. Customized scripts and Oracle forms are used to manage and store metadata. The APFO investigated the use of a commercial metadata management tool from Compusult Ltd., but has put those efforts on hold.

Telecommunications

The APFO has one 56-kilobyte (KB) line operational and a T1 line installed to support WAN communications. Representatives from the Electronic Access Initiative (EAI) Team visited the APFO in FY01 to perform a suitability survey for the installation of the OCI Security Stack and to assess WAN requirements. Internally, the APFO is transitioning to a gigabit network to handle large digital imagery data sets. This upgrade is ongoing and did not serve all production units at the time of this report.

NCGC – Fort Worth, TX

The National Cartography and Geospatial Center (NCGC), Natural Resource Conservation Service is a technical leader for NRCS in cartography, remote sensing, global positioning systems, and Geographic Information Systems (GIS). NCGC is a clearinghouse for geospatial data sets and the preparation of maps and map products. NCGC provides innovative leadership in the map-making process. Generally, NCGC supports the soil survey process by providing map products and materials necessary for the completion of soil survey publications.¹⁶

Hardware

NCGC supports a mix of hardware platforms including Macintosh, Unix, and NT based platforms. The majority of desktops supporting geospatial production are NT bases, whereas the storage is primarily a Sun configurations. There are some exceptions to this generalization. NCGC uses tape storage devices primarily for backup and archive. On-line storage is supported by RAID arrays running on Sun Enterprise class servers.

Software

NCGC cartographic and geospatial processing production is primarily performed using the ESRI suite of products including ArcView, ArcView Extensions, ArcInfo, and SDE. Some support to data dissemination is supported by the ESRI Web Mapping (ArcIMS) and OGC Web Mapping technologies. NCGC is transitioning from Informix to Oracle for most of their RDBMS requirements. NCGC had plans to evaluate a commercial metadata management tool from CompuSult Ltd., but has put that effort on hold. NRCS ITC has migrated some of the “back-end” components of the Resource Data Gateway to NCGC. This system includes a scripts and services designed around the ESRI ArcIMS software.

Telecommunications

The internal network at NCGC is composed of Cisco® equipment (routers, switches, and a firewall) configured to secure and deliver 100-megabit capability to each desktop. Externally, the NCGC WAN connectivity include two T1 lines dedicated to center operations and one other T1 shared with other NRCS components located at the Ft. Worth site.

Security

NCGC has begun the transition of some equipment to support the recommended OCI Security Stack.

Data

NCGC currently maintains XXX TB of on-line and archived data. The majority of this data is Digital Ortho Quad (DOQ) imagery in support of the soils digitizing centers and

the collarless Digital Raster Graph (DRG) database. NCGC manages a wide variety of vector products as well, most notably the SSURGO database. Much of the data produced by NCGC is ordered on-line and sent via File Transfer Protocol (FTP) or Compact Disc Read Only Memory (CDROM). NCGC is piloting the use of on-line data access through various web-mapping methods including ArcIMS and the Massachusetts Institute of Technology (MIT) Ortho Server. Metadata for these products is file based and is served to the public through their NCGC NSDI node.

Staff

NCGC currently employs approximately six full-time IT support staff to assist approximately 90 production staff. The IT support staff provides hardware, software and network support to the production and data dissemination activities. Their skills include database administration, web site administration, network support, and hardware maintenance. NCGC staff provides secondary Help Desk support to service centers, states and the public. NCGC has three certified ESRI software trainers on staff.

NWCC – Oregon

The National Water and Climate Center (NWCC), has the lead in NRCS to acquire, develop and transfer water and climate information and technology to the agency and its partners to support natural resource conservation. The NWCC houses data acquired from NWS' COOP network, as well as data acquired with the agency's own data collection networks, SNOTEL (Snow Telemetry) and SCAN (Soil Climate Analysis Network). Both of these networks are comprised of automated remote sites, which transmit data as frequently as hourly to the Centralized Computer Facility in Portland, OR. This facility and the associated data collection and dissemination are a 7/24 operation, with minimal downtime allowed. The center supports data management, data analysis and product generation. The Center is also responsible for developing and acquiring other water resource related technology for use by NRCS and its partners and cooperators

Hardware

NWCC supports a mix of Intel and RISC based systems, running Windows NT, as well as UNIX systems. The desktop environment is Windows based, while the Data management, storage, distribution and some of the analysis are performed on UNIX systems. The data management and analysis is located on systems running HP-UX, while data distribution (FTP and Web-based), are served by SUN Solaris. Central servers are backed up using tape drives. Operational storage is done using a combination of regular hard disk storage as well as RAID storage. Currently the data collection, quality control, and data management is done on one HP K series server, while data analysis and product production is done on HP I series. Both of these servers, as well as test and development systems are behind an authentication server running Windows 2000 Advanced Server. A web and ftp server operates on a SUN system running Solaris 8.

Software

Software for these systems include Informix DBMS, along with custom designed applications. The SNOTEL system, which uses Informix for data and metadata, custom software application written primarily in C, C++, and Java, provides the custom software for data management, quality control, and dissemination of data to customers and web servers. The Centralized Database System (CDBS) and Centralized Forecasting System (CFS) systems are based on custom databases and associated custom software applications. These applications are migrating to a NetCDF file format to improve both performance and portability of these primarily time-series datasets. Currently the custom software is a mix of Fortran, C, C++, Java, and other scripting languages. Some of the analysis tools for the CFS system are Personal Computer (PC) based custom applications.

Telecommunications

The internal network at NWCC is composed of Cisco equipment (routers, switches, and hubs) configured for 10-megabit capability. The NWCC WAN uses a T1 line to an OCIO stack located several blocks away, with access to both the USDA backbone and the Internet. This T1 is shared with other USDA agencies located in Portland (NRCS State Office, Rural Development, and FSA).

Security

NWCC is planning a security architecture compatible with that recommended for use with the web farm infrastructure. In addition, NWCC is in the process of implementing a Public Key Infrastructure (PKI) solution for secure access to the NWCC servers associated with data analysis and product generation. This is a pilot project, but it is expected to be maintained beyond the period of the pilot.

Data

NWCC currently maintains approximately 10 GB of on-line spatial data layers from their Parameter-elevation Regressions on Independent Slopes Model (PRISM) and approximately 70 GB of transactional data to support programs including SNOTEL, CDBS, AWDB, WYSNO and WYFOR. The NWCC is migrating their CDBS, WYSNO, and WYFOR applications to AWDB. This is a data management system based on a NetCDF file structure, with Informix as the metadata database. The NetCDF structure is optimized for use with time-series data provides rapid access to data. The NWCC is also in the process of augmenting the dissemination of their information, and their access to climate information and applications through the Unified Climate Access Network (UCAN) program. UCAN will provide users with "virtual access" via the Internet to climate datasets collected by federal, state, and county networks. UCAN will provide a networked computer and data storage infrastructure that will allow users to access climate more efficiently than previously possible. Users will be able to access climate applications as well.

Staff

NWCC currently employs seven full-time IT support staff to assist the NWCC staff and the SNOTEL data acquisition system. The IT support staff is organized into Operations, which handles server system administration, telecommunications, networking, web-site administration, SNOTEL system operations, and desktop configuration and support; Data Management which is responsible for database management for the SNOTEL, SCAN, and Climate databases; Software Management, which handles custom software maintenance and software development.

Current CCE Architectures and Impacts

Intermediate Data Architecture

The intermediate data architecture calls for static and dynamic GIS data sets to be stored at local Service Centers. Changes and updates made to dynamic data sets, such as CLU, will occur locally and be replicated to the regional and or national level. Service Center users will likely access and utilize geospatial data sets at the local level while other agencies and the public will rely on geospatial data warehouses at the national level data centers.

This architecture forces day-to-day GIS system demand down to the local level and has the added advantage of decreasing the load on the USDA communications network. Additionally, this supports the levels of reliability and performance that USDA requires for doing business. However, this architecture does present two risks, one concerns the state of replication technology and the second concerns an increased need for additional administration at the Service Center.

The technology used for geospatial data replication using ESRI's ArcSDE is not projected for release until mid-2002, which limits the ability to manage dynamic data sets and take advantage of versioning and history tracking. The intermediate architecture also requires a much higher degree of administration due to the maintenance of a RDBMS and the GIS databases managed by them.

Long-term Data Architecture

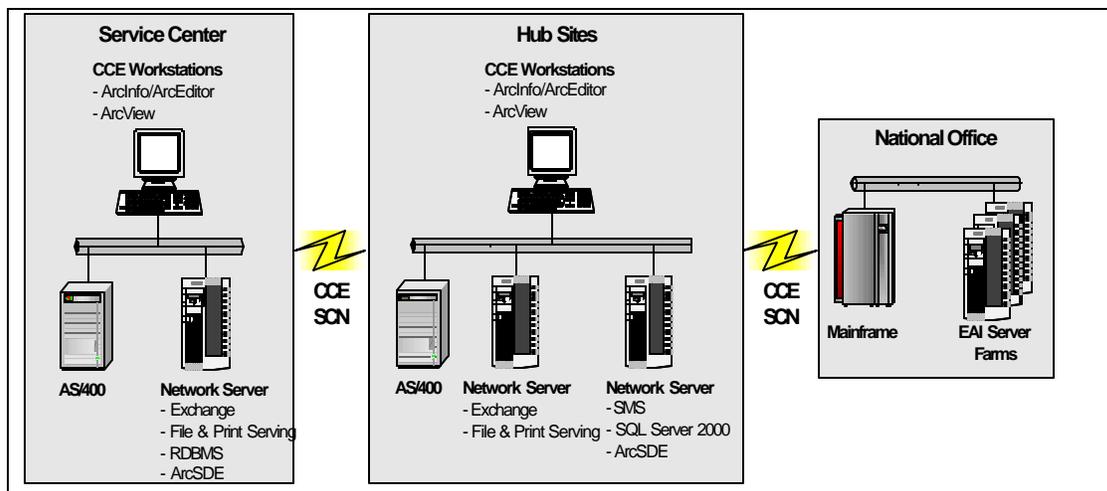
The long-term data architecture is geared towards supporting growth in the area of Internet GIS, which requires a much higher bandwidth than currently available at Service Centers. As technology and bandwidth increases are realized, certain USDA GIS applications will migrate to a Web-based centralized application and data architectures, resulting in lower administration and maintenance costs.

CCE predicts that adoption of this long-term architecture will vary by agency and some agencies will sustain characteristics of both the intermediate, or local data storage and application hosting and long-term where a combination of local and Web-based data and applications.

Recommended Technical Architecture

The ITA identifies an enterprise-level Technical Architecture for the three Service Center models, local, hub and national. The technical architecture defines the hardware and software that comprise each architecture and how the models connect through the Service Center network. After reviewing central and distributed models for application hosting, CCE recommended a hybrid architecture based on upgrades to the existing telecommunications infrastructure and utilization of existing server capacity. CCE found that small Service Centers are well positioned to handle the additional load geospatial software would place on their existing hardware configuration. Larger service centers and hub sites can initially leverage excess server capacity until two new network servers are deployed to each hub in January 2002. The addition of these servers will allow hubs to manage the increased capacity of geospatial processing.

The hybrid architecture addresses the intermediate-term data architecture and the local data and application requirements, while preparing for the long-term data architecture that moves more applications and data towards Web-based central distribution. This architecture configuration is based on current telecommunications limitations and the premature nature of some Web-based GIS technologies that are required by USDA Service Centers. This architecture will serve USDA well until a transition towards a more Internet-based business model is phased in as technological advances are realized. The hybrid architecture is presented in Appendix A: Figure 0-1.

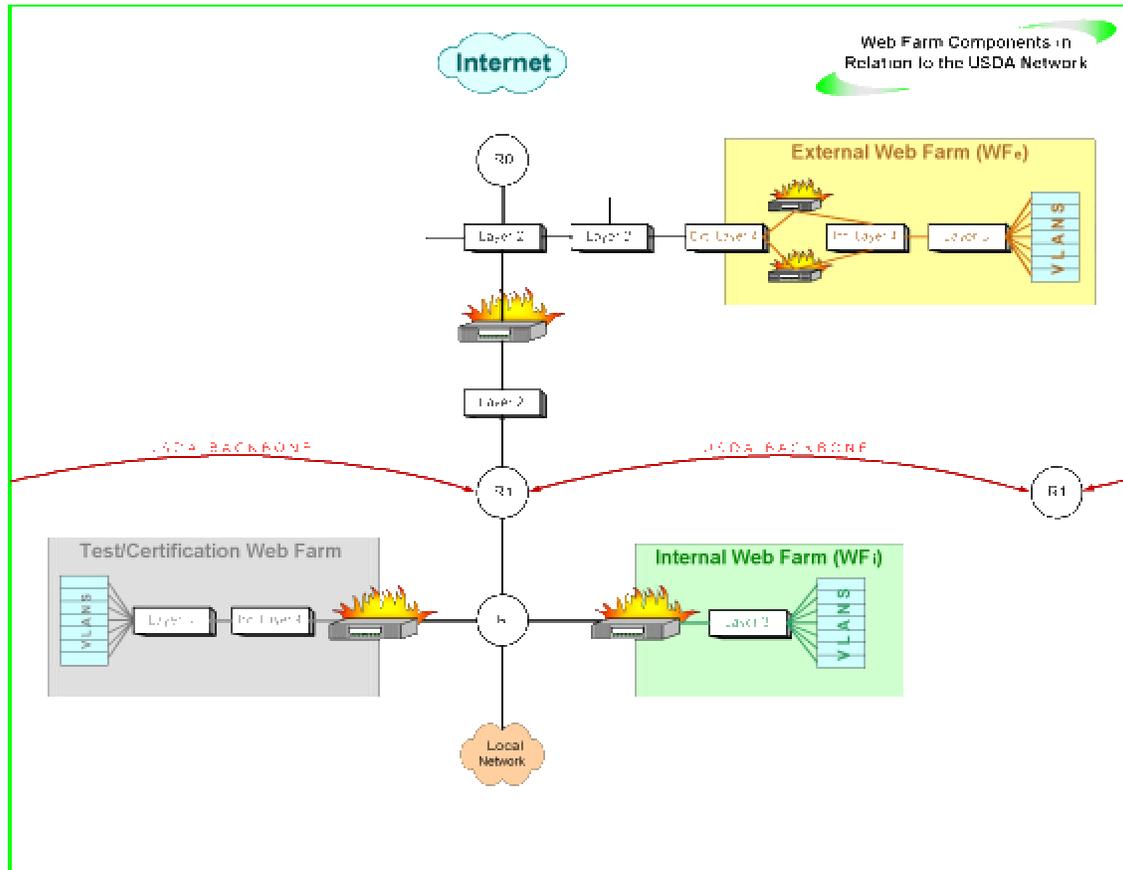


Appendix A: Figure 0-1 Sample Hybrid Application Architecture

This hybrid architecture approach is consistent with the architecture alternative example presented in Figure 4-2. It demonstrates the current state of a “virtual” or point-to-point data warehousing structure represented at Service Centers and hubs and the ultimate migration towards the distributed data warehouse architecture where users are accessing a series of data marts established by ownership primarily along agency lines at the national level.

Current EAI Architecture and Impacts

The current configuration of the EAI Web Farms in relation to the USDA network is shown in Appendix A: Figure 0-1



Appendix A: Figure 0-1 Web Farm Components in Relation to the USDA Network

OCIO Stack(s)

The OCIO Stack is a term used to describe an identical configuration of network devices at each of the seven gateways from the Internet (UUNET) to the Internal USDA Network. The OCIO consists of the following:

1. R0 – router
2. Sx0 – layer 2 switch required for the Intrusion Detection network probe
3. F/W - firewall
4. Si1 – layer 2 switch
5. R1 – router. Connections (T-1s) between R1 routers at each of the seven gateways that make up the USDA Backbone.

External Web Farm (WFe)

The purpose of the External Web Farm (WFe) is physical and network containment of Web-based services (Intranet, Extranet, and Internet) for clients independent of source network. Authorization to services is based solely on user authentication rather than source Internet Protocol (IP) address. The WFe is primarily designed to house interface-layer applications. Middle-tier and backend applications can be placed inside the WFe (in VLAN 5 and 6 respectively) if these applications are not shared externally. In other words, the only way to reach a non-interface application resident in the WFe is through an interface-layer also resident in the WFe. The WFe provides for both local and global load balancing across the three EAI WFe's (FC/KC/STL).

The WFe has nine discrete VLANs:

- VLAN 2 - Anonymous Public Services
- VLAN 3 - Internal Services (Domain Controllers, Policy Servers)
- VLAN 4 - Non Conforming applications
- VLAN 5 - Isolated Middle Tier Service (self-contained in WFe)
- VLAN 6 - Isolated Database and Backend Service (self-contained in WFe)
- VLAN 7 - Secure Farm Mgmt (Mgmt Consoles, Internal DNS, CVS, etc)
- VLAN 8 - Authenticated Public Services
- VLAN 9 - Reserved for Future Use
- VLAN 10 - External Services (SMTP Gateway, External DNS)

Internal Web Farm (WFi)

The purpose of the Internal Web Farm (WFi) is to provide physical and network containment of services (both Web-based and client-server) for clients resident on the USDA Internal Network. The WFi can house *Intranet-only* interface-layer applications, although the WFe provides considerably better scalability and extensibility than the WFi for these applications. The WFi was primarily designed to house middle-tier and backend services that are shared with Web Farm interface-layer applications and other non-Web Farm applications. Examples would include a database that requires ODBC access from other non-Web Farm databases, or other interface-layer applications. An interface-layer application running in the WFe can connect to a backend service anywhere on the USDA internal network. The main advantage of the WFi is that the EAI Infrastructure is shared across both the WFe and WFi environments. This allows the storage area network, backup network, and security perimeter to be extended to the internal network.

The WFi has five discrete VLANs:

- VLAN 2 - Anonymous Intranet Services
- VLAN 3 - Web Farm Internal Services (AgWeb, Policy Servers)
- VLAN 4 – Non-Conforming Applications
- VLAN 5 –Middle Tier Services
- VLAN 6 –Database and Backend services

Test/Certification Web Farm (WFtc)

The purpose of the Test/Certification Web Farm (WFtc) is to replicate the WFe and WFi environment for the purpose of testing and certifying applications for hosting, configuration changes, patches, and upgrades of web farm and server components in a safe environment. This environment is made necessary by both the need to certify applications prior to be hosting in the Web Farm as well as the complexity of the Web Farm environment.

Staffing

The Web Farms are staffed with Web Administrators, who have primarily system administration-type duties and expertise. The existing staffing levels are adequate for the current workload, but would be insufficient for additional workload required for the GIS Data Warehousing project. There are no GIS or database specialists assigned to the Web Farms; although these resources can be obtained on a limited, prearranged basis as needed for individual projects, provided funding is available.

Bandwidth

The current bandwidth between the internal and external Web Farms (see Figure 9-1) is approximately 100 megabits per second (Mbps). This capacity is sufficient for the current applications being hosted in the Web Farms, but not adequate for future applications that entail large or frequent transactions, such as GIS data marts. The external bandwidth depends on the USDA backbone; again, the capacity is sufficient for current applications, but not adequate for future applications that entail large or frequent transactions.

Appendix B – Replication Options - Gartner Group Report

Platform Dependency	Company/Product	Supported Host Platforms	Transaction Aware?	Replication Type: Mirroring or Shadowing	Replication Selection Granularity	Production Deployment to 100+ Customers?	Simultaneous Use of Replica for Reporting?
Host-independent, disk-dependent (disk-to-disk remote copy)	Compaq DRM	Unix (Compaq, IBM, Sun), Windows	No	Mirroring	Physical disk volume	No	No
Host-independent, disk-dependent (disk-to-disk remote copy)	EMC SRDF	MVS-OS/390, Unix (IBM, HP, Sun), Windows, Unisys, Bull, FSC, ICL, AS400	No	Mirroring	Physical disk volume	Yes	No
Host-independent, disk-dependent (disk-to-disk remote copy)	Hitachi HRC	MVS-OS/390, Major Unix (HP, Sun, IBM AIX and NUMA-Q, Compaq, NCR, SGJ), Windows, NetWare	No	Mirroring	Physical disk volume	Yes	No
Host-independent, disk-dependent (disk-to-disk remote copy)	Hitachi HOARC	MVS-OS/390, Major Unix (HP, Sun), Windows	No	Shadowing	Physical disk volume	No	No
Host-independent, disk-dependent (disk-to-disk remote copy)	IBM PPRC	MVS-OS/390, Unix (IBM AIX and NUMA-Q, Sun, HP, Compaq, DG), Windows, Novell NetWare	No	Mirroring	Physical disk volume	Yes, OS/390 only	No
Host-independent, disk-dependent (disk-to-disk remote copy)	HP Continuous Access XP	Major Unix (HP, Sun, IBM, Compaq), HP MPE, Novell, Windows and Red Hat Linux	No	Mirroring	Physical disk volume	No	No
Host-independent, disk-dependent (disk-to-disk remote copy)	HP Continuous Access XP Extended	Major Unix (HP, Sun, IBM, Compaq), HP MPE, Novell, Windows and Red Hat Linux	No	Shadowing	Physical disk volume	No	No
MVS-OS/390	ENET RRDF	MVS-OS/390	Yes; supports DB2, IMS, CICS, IDMS, CPCS, ADABAS and SuperMICR	Electronic journaling (logging) and optional shadowing for IMS and DB2	DBMS (IMS and DB2)	Yes	Yes
MVS-OS/390	Amdahl XRC, Hitachi HXRC, IBM XRC	MVS-OS/390	No	Shadowing	Physical disk volume	No	No
NSK (Compaq Himalaya)	Compaq RDF	NSK	Yes	Shadowing	Transaction	Yes	Yes
Oracle	Oracle Standby Database	All platforms Oracle DBMS supports	Yes	Shadowing; log-shipping and apply	Database	Yes	No
Oracle	Quest Software SharePlex (for Unix)	Unix (HP, Sun)	Yes	Shadowing; log-based	Tablespace	Yes	Yes
OS/400	Data Mirror High Availability Suite	OS/400	Only when using AS/400 commitment control	Shadowing; wide-area cluster support	DB2/400 DBMS, OS/400 object	Yes	Yes
OS/400	Lakeview Technology MIMIX software suite	OS/400	Only when using AS/400 commitment control	Shadowing; wide-area cluster support	DB2/400 DBMS, OS/400 object	Yes	Yes
OS/400	Vision Solutions High Availability Suite	OS/400	Only when using AS/400 commitment control	Shadowing; wide-area cluster support	DB2/400 DBMS, OS/400 object	Yes	Yes

Appendix C – Bibliography

¹ United States Department of Agriculture. *Phase II: Implementation of Geospatial Data Warehouses*. September 2000.

² United States Department of Agriculture. *Requirements and Cost Benefit Analysis. Managing Geospatial Data for Better Program Delivery. A Service Center Initiative*. June 22, 2001.

³ United States Department of Agriculture. *Food and Agricultural Policy: Taking Stock for the New Century*. September 19, 2001.

⁴ Federal Geographic Data Committee, *Geospatial One-Stop, Office of Management and Budget Capital Asset Plan*, February 4, 2002.

⁵ Inmon, W.H. *What is a Data Warehouse?* Volume 1 No. 1 Prism Solutions, Inc. 1995

⁶ Ian Manning, *Data Warehousing - What Is It?*

⁷ White C., *Managing Distributed Data Warehouse Metadata*, Data Management Review, February 1999

⁸ Inmon, W.H., *OLAP and Data Warehouse*, 2000.

⁹ Inmon, W.H., Zachman, J.A., Geiger, J.G., *Data Stores Data Warehousing and the Zachman Framework*, 1997.

¹⁰ Orr, K., *Data Warehousing Technology*, 2000.

¹¹ National Cartography & Geospatial Center Annual Summary — FY 2000.

¹² United States Federal Geographic Data Committee, *A Strategy for the NSDI*, April 1997.

¹³ Walker, R.W., *Patience is a virtue when building data warehouses*, Government Computer News, February 22, 1999.

¹⁴ Gartner Research Group, *Disaster Recovery: Weighing Data Replication Alternatives*, 15 June 2001.

¹⁵ Inmon, W.H., *Information Management: Charting the Course: Iterative Development*, Data Management Review, May 1999.

¹⁶ National Cartography & Geospatial Center Annual Summary — FY 2000.