



The Data Architecture

Planning the foundations

DRAFT

Version 1.5

February 1, 2000

Purpose

The information needs of an organization are constantly evolving, as business strategies evolve and the business environment changes. Because of this, the technological constructs deployed must be able to evolve along with the business, in order to maintain the ability to provide useful and effective information to business people. The key to providing this evolutionary capability is a well defined architecture.

An information system is more than a collection of tools and technology. Architecture must specify the conceptual layout of the data processing environment, the hardware and software that support that environment, and the standards and procedures that make it function. It is crucial that this architecture is defined before hardware, software, and methodology decisions are made. All too often organizations select a component based on some perceived valuable attribute of that component, only to find out that it is very difficult to integrate that piece of technology into their overall architecture. We must first identify the logical components and functions of the warehouse architecture, and then select the best hardware and software that fits into this architecture.

The purpose of this document is to provide a set of strategic data and technical architectural alternatives for The United States Department of Agriculture's Service Center Initiative Team and its customer agencies (FSA, NRCS, and RD). The alternatives proposed encourage the convergence to a common architectural framework that promotes heterogeneity, interoperability, and extensibility that will serve as the foundations for the continued growth and effectiveness of the USDA's modernization plan.

Although some sections of this document discusses specific hardware and software product brands, this document and the strategies contained herein, make no recommendation as to 'brand-name' hardware, or commercial off the shelf products. In fact, product brand names will be dutifully avoided as such recommendations are beyond the scope of this discussion. Additionally, this document does not preclude solutions selected by BPR teams, specific solutions to specific agency problems or any other development activities underway. Further, this document does not mandate an immediate migration to any specific systems environment or technology.

Rather, this paper addresses the architectural constructs required to support the business goals and information needs of several hundred users across a national service area footprint. Principally addressing the Data Architecture, this document provide a template/blueprint for several data architectural alternatives and a context for planning, budget formulation, and investments in information technology by the USDA.

Background

The United States Department of Agriculture agencies, including the Natural Resources Conservation Service (NRCS), Farm Service Agency (FSA), and Rural Development (RD), recognize the need for timely access to comprehensive, accurate, timely and relevant information on which to base important business decisions. It is more important today than ever for business personnel to have timely, efficient access to data that is currently available only through disparate legacy systems. The implementation of an integrated data architecture will significantly improve the information availability in support of a wide range of loan, farm and conservation programs.

The Service Center Initiative (SCI) intends to bring about these improvements by reengineering the processes involved with the management, delivery and access to business information. SCI recognizes that common standards, languages and systems that support data processing can be integrated and shared across the partner agencies and are key ingredients in improving delivery of information and services to USDA customers. This initiative seeks to leverage these opportunities for improvement, along with advances in systems and telecommunications technology to significantly impact how the partner agencies work together to carry out their missions.

Service Center Initiative's Basic Goals and Objectives

The Service Center mission statement describes a single center that offers the products and services of the three Partner Agencies in a manner that provides customers the best possible service at the least possible cost. The four Service Center objectives, described below, elaborate on that vision¹.

One-stop shopping

One-stop shopping means that agricultural, rural development, and natural resource conservation programs are provided to customers in a timely, competent, and thorough manner by Service Center employees without regard to the responsible Agency. The Service Center Partner Agencies will offer exceptional service—seamlessly, as if the Partner Agencies were one.

Quality customer service

Quality customer service is defined as providing courteous, high-quality, professional, and personalized service in a timely and effective manner that exceeds customer expectations.

Cost reduction

Cost reduction entails an ongoing effort by USDA Service Centers to reduce administrative and program delivery costs to the public by utilizing integrated information systems, and sharing administrative resources to the maximum extent possible.

¹ Adapted from the USDA Service Center Business Need and Technical Evaluation Study, November 28, 1997, p 2-9 through 2-10.

Partnerships

Partnerships between USDA Service Centers and people, communities, and other private organizations and government Agencies will maximize the use of limited resources and help all partners attain their goals and objectives.

The Role of the Service Center

The role of the Service Center is to be a new delivery mechanism for the programs, products and services offered by the Partner Agencies. This delivery mechanism will use redesigned work processes and integrated systems and data to provide the best possible service at the least possible cost.

FSA, NRCS, and RD are in the process of collocating field offices into local Service Centers. Although these Agencies served the same regions, worked with overlapping groups of customers and performed many similar processes, they operated as separate entities with their own operations, data, information systems and service delivery processes. The creation of the Service Centers is intended to improve the service delivery to USDA customers and capture cost savings from both collocation and rationalization of processes and systems across the Partner Agencies. However, until current BPR and IS initiatives are completed and implemented, many Service Center operations will continue to resemble their historical operations.

Roles of the Partner USDA Agencies

Farm Service Agency (FSA)

Helps American agriculture with commodity, credit, export, and risk management programs intended to improve the economic stability of agriculture. These programs help keep enough farmers in business to produce an adequate food supply and to keep consumer prices reasonable.

Natural Resources Conservation Service (NRCS)

Works with landowners to develop conservation systems uniquely suited to their land and individual ways of doing business. NRCS also provides technical assistance to communities to help solve resource problems, and protects soil, water, air, plant, and animal resources to meet the needs of this generation without compromising the welfare of future generations.

Rural Development (RD)

Helps ensure that rural citizens can participate fully in the global economy by providing technical assistance and programs that help rural Americans build strong local economies. Helps rural communities meet their basic needs by financing water and waste water systems; financing decent, affordable housing; supporting electric utilities and rural businesses; supporting community development with information and technical assistance; and providing emergency disaster assistance and relief.

These three agencies together provide a variety of programs and services to a wide and often overlapping clientele base.

Although the Partner Agencies share much of the same clientele and have some common processes, there are basic differences in how they work and where they interact with customers. NRCS is county-based with a dispersed, highly mobile workforce. It relies on many partners to fulfill its mission, chief among them the 3000 local conservation districts that link NRCS to local priorities for soil and water conservation.ⁱ

FSA is also a dispersed agency with a “grassroots” delivery system. However, because of the nature of many of its programs, individual customers tend to come into the local FSA office to apply for benefits and to check on application status.

In contrast, many of RD’s operations involve partnerships and cooperation with local community governments, non-profit corporations, and lending institutions, instead of with individuals.ⁱⁱ RD has locally-based staff, but as of December 1996, it had personnel in only slightly more than half as many Service Center sites as either FSA or NRCS.ⁱⁱⁱ

The Emerging Architectural Vision[§]

Recent performance and simulation modeling of various architectural alternatives have confirmed the need for management of some spatial and tabular data at the Service Center level of the architecture. This is based upon the recognition that the LAN/WAN/Voice infrastructure cannot support the transfer of the volume of data associated with the Enterprise GIS solution. In addition, BPR projects have presented additional requirements that support the need for local application and data services. The CCE Team is still in the process of evaluating various architectural options to determine specific placement and sizing of local servers.

The current SCI Modernization Plan^{iv} outlines an architectural vision based on the premise that limited investment funds, together with changes in technology creates the opportunity to investigate an option that focuses on providing applications and database servers to key “core offices”.

This new architecture proposal focuses on identifying a smaller number of Service Centers as “core offices” or centers of investment that will house the information technology (IT) infrastructure. The IT infrastructure will be accessible to all offices through the USDA telecommunications network. Initial estimates envision 800 to 1,200 core offices.

Non-core offices will receive the same services as core offices, without the necessity to house and manage an increasingly complex server environment. Non-core offices will exist where there is a program delivery or economic justification to do so. Non-core offices will receive little or no technology investments beyond personal workstations, printers, and telecommunications connectivity—thereby significantly reducing the overhead cost of smaller, non-core offices. Limited IT staff and other specialists can be used more effectively in supporting an infrastructure concentrated in fewer locations. Depending upon technical and business requirements, various components may be housed nationally, regionally, at the state level, or on a multi-county basis. The CCE

Team currently is investigating the feasibility of this option through the use of performance modeling and simulation

In terms of the server platform, the *USDA Service Center Business Need and Technical Alternative Evaluation Study – Phase II, April 9, 1998* identified 19 options for application server configurations. These were later reduced to the following three options:

- Windows NT Server - Local level
- Unix Server – State/Regional Level
- AS/400 – State/Regional Level

Of the three applications server solutions under consideration from that study, the solution presented in this initial document (Windows NT Server @ Local Level) represents the only remaining applications server solution that provides a server implementation at the Service Center level. The hardware platform proposed is the one currently being piloted in the nine BPR pilot sites, and so represents the solution that has been proven to support reengineered applications in a production environment. At this time, the intent is to focus on the known technology needs and provide an initial perspective on how those needs may be met. There is no intent to eliminate any of the other possible solutions for consideration as part of the final CCE. In fact, it is believed that all three of the remaining application server options will be part of the final solution.



USDA Service Center Initiative Pilot Service Center Locations

Currently, the architecture of the pilot sites is limited to a series of local stand-alone databases. This architecture does not support application integration, data warehouse

initiatives, or direct connectivity to legacy data. The storage requirements at the local level are large and the ability to maintain one hundred percent of the data at the local service center would be costly and challenging. The alternative that follows, as well as the CCE investment center alternative described above, assumes some level of data storage at the local service center. This is based on the premise that the closer the data is to the user, the smaller the amount of data that must travel through the data links to reach the user, thereby decreasing user wait time.

Extending the Vision

The Data Architecture is one of the fundamental components of the enterprise technology architecture. Most developers are familiar with the application architecture's data design determining what data is needed for a particular application. At an enterprise level however, we need to look at how data is managed, accessed, and stored in databases across multiple applications in the organization. The data infrastructure must be rigorously architected, which is no small feat by any measure in large organizations, and must be conducted by skilled, experienced architects under the guidance of business-area managers and in corporation with information management teams and functions.

Architectural Context

Architecting the data architecture requires the categorization, modeling, and evaluation of an organization's data. Once this is done, the physical components necessary to store, transport, view, and analyze the data can be analyzed and defined. With this comes the challenge of management, both of systems and of data, and the integration of disparate systems to gain value from each of the data domains within the organization.

To accomplish this, a pair of concurrent, related efforts must be undertaken. The first effort is centered in identifying the models and principles that define the *use and positioning* of data. The second is the evaluation and identification of the technologies that will support the data environment.

Methodically Approaching the Physical Data Architecture Design

Although it is beyond the scope of this document to delve into the details of the varied methodologies for developing Information System Architectures, some time will be spent in the following section to provide an overview of the activities involved in arriving at solutions for enterprise architectural designs. The distinctions between the physical and logical architectures aren't artificial. Putting together these different architectures means considering different issues, and it requires two completely different sets of tools and skills. Following are some guiding principles, arranged in the order of life-cycle execution.

Beginning with the *softer* analytical tasks, three tasks are involved. Following is a brief description of the tasks and they are...

- Information Synthesis;
- A Data Usage Model; and
- A Data Distribution Model

Information synthesis examines how data is combined, aggregated, derived, and disseminated to provide useful information.

The Data Usage Model is a set of descriptions that surround each of the individual uses of data within the organization. The data usage mode is itself divided into two sections: the "*nature*" of the data (analysis oriented, decision support, management, reporting, transaction processing, etc.) and the "*orientation*" of the systems that support such usage, including transactional systems, operational data stores, data marts, or data warehouses.

The Data Distribution Model focuses on the positioning of data within the enterprise and the relative proximity of such data to its sources and its uses.

Upon completion of these efforts, (and not before) the relevant technologies can be aligned to the needs of the organization. Given the nature of the data, its uses and distribution, different technologies can now be combined to provide real value to the end user. Technology is useful only if it supports the performance, access, and functional needs of the end user.

At this stage, evaluation of commercial off the shelf and third party technology such as the relational database management systems (RDBMS), multidimensional database systems, data access tools, data transformation technology, transaction monitors, online analytical processing tools (OLAP), and each of their related hardware platforms can be conducted for need and fit. Simplicity is the rule here. The more technology added to the mix, the more difficult it will be to support.

Next, good practice requires the organization to develop and compile standards that will be used throughout the remaining phases of the life-cycle. These standards, cover all aspects of data, not just the software used to access it. Standards related to the physical architecture cover the hardware platform, database management system(s), access methods, abstraction methods, programming model, coding principles, interface agreements, data management processes, and technology choices. These standards are used to achieve two major things: they provide a consistent approach to systems development, which helps the organization maintain a single orientation over systems, and they form the basis for technology selection. In conjunction with these standards, an understanding of the volumes and activity around the data must be gained.

Volometrics aren't limited to the number of rows or data elements or table sizes associated with a given schema; they include the following information:

- Interface processing volumes
- Interface processing windows
- User population data access characteristics (reads and writes)
- Table usage (hot tables vs. cold tables)
- Table growth rates
- Index growth rates
- Table row populations

- Table row sizes
- Database page/block sizes
- Operating system block size

These are quantifiable datum values that lend themselves to the appropriate sizing of the hardware needed to support the data in an enterprise scale architecture. More importantly, the understanding of the activity of the data, supports the application of the *proper* technologies for the *proper* purposes.

The use of volumetrics early in the process is generally an exercise in estimation, based on the data architect's experience, and the enterprise's knowledge of its data at the current point of time. As the process of designing the architecture proceeds, the volumetric assessment of the enterprise's data will become more the result of access patterns and enterprise experience.

With the standards, volumetrics and the softer analytical tasks completed and in place, an understanding of the proposed uses of the data is readily available, and through use of the distribution model, usage model, and the volumetrics associated with the data, recommendations and adjustments to make the system perform (network, server, data storage requirements, etc.) can be made with confidence. Equally important, it is that it is now possible to apply the technologies (hardware and software) such that they solve real business problems and exist not only as a set of aggregated tools and technologies.

On this foundation, the next (and most recognized) components of the data architecture are built--the schemas and interfaces. *Schemas* are created from three major components: The logical view of the data usually in the form of a entity-relationship diagram, metadata definitions and models, and the enterprise logical data model. Assuming these views of the data are in a normalized state, the next and final task is to denormalize the models/data, create a *denormalization map*, and define the interfaces that get or provide data to or from other systems.

The Denormalization Map will outline each of the denormalization steps taken to improve performance and provide useful data to the end user. An accomplished denormalization mapping will not only discuss the individual schema mappings, but will also cover such features across schemas. The denormalization mapping feeds into the metadata repository, and helps the enterprise prevent the loss of knowledge about how its systems derive information.

Summary

Increasingly, system development efforts have not lived up to the requirements or expectations of their users or their procuring agencies. This lack of success is symptomatic of a larger problem, which is rarely traceable to the hardware or software technologies, but instead lies in the fundamental methods used to engineer systems. The approach presented here has proven successful in constructing architectures equally as large as that of the USDA's.

To-date, most of the tasks described above have not been executed, and those which have, have been done so outside of the framework of a recognizable methodology. To accomplish the goals set out by the USDA's Modernization Plan and the Service Center Initiative, SCI's management must resolve to a proven, methodical approach must be adopted to ensure the viability of technologies deployed and investments undertaken.

Open, Heterogeneous, Distributed Processing Architectures

Open Systems and Heterogeneity

Any modern, viable architectural framework relies on the concept of “*open-ness*” for successful implementation. The objective of open systems is to help users procure, and vendors provide, systems that meet user requirements for information technology functionality and at the same time enable users and sponsoring agencies to...

- Protect their investment in information technology in a rapidly changing world
- Adapt and evolve their systems as business needs and technologies change
- Use systems and software from many suppliers (e.g. Allow for the establishment of a heterogeneous environment)

The IEEE defines an *open system* as a comprehensive set of interfaces, services, and supporting formats, plus user aspects for interoperability and for portability of applications, data, or people, as specified by information technology standards and profiles. Similarly, the X/Open Consortium defines an open system to be a vendor and technology independent, interoperable computer environment comprised of commonly available products, implemented using accepted methods and de facto standards. It implements sufficient open specifications for interfaces, services, and supporting formats to enable properly engineered application software...

- To be ported with minimal changes across a wide range of systems
- To interoperate with other applications on local and remote systems
- To interact with users in a style that facilitates user portability

Additional context for “open systems” criteria are found in definitions by some of the industries' renowned experts. From Hewlett-Packard: “Open systems are software environments consisting of products and technologies which are designed and implemented in accordance with standards—established and de facto—that are vendor independent and commonly available.” From Gartner Group: “An open system is an information processing environment based upon application programming interfaces and communication protocol standards which transcend any one vendor and have significant enough vendor product investment to assure both multiple platform availability and industry sustainability.”

In short, open systems are heterogeneous by nature and compliant with industry-recognized standards. For the architectures proposed herein, this is an essential and fundamental concept.

*Finally, the most important consideration for open, distributed architectures is the network. The network is the **key** to successfully deploying open systems. It is the integrating and communicating force in the environment, providing access to multiple services and carrying several protocols in support of these services. The importance of the network cannot be understated. Open systems are network-based and workstation-centered environments. It is imperative that any initiative to move to open systems also carries with it initiatives to upgrade network infrastructure designs, complete with support for upgrades of standard workstations and local area networks.*

Distributed Processing Architectures

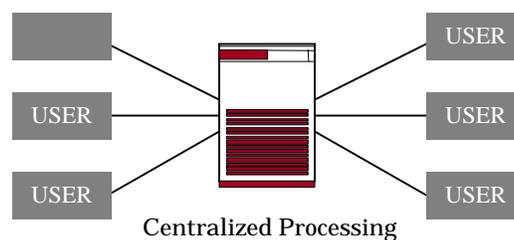
The hardware design issues are centered on the model selected to manage data processing. The node or nodes within the architecture with the fastest or most powerful computing power are assigned the role of server. As server, it handles the processing in some cases of the user interface, application execution and data storage. The two important aspects to be considered in choosing server nodes are processing power and storage capacity.

The processing power of a server is critical to the response time available to queries. Another concern related to the server's processing power is the fact that other processes may be competing for processing time other than data base requests. Therefore, an excessive amount of processing traffic on a server can cause monumental performance problems in the distributed data base environment. As compared with the server node, the client node can be limited in its amount of processing power and storage capacity. Client nodes are typically smaller desktop microcomputers or workstations.

Four basic types of distributing processing architectures exist – peer to peer, client server, three or 'N' tier, and cooperative processing. The physical orientation for each of these is depicted in the figures below. Although the figures shows a fairly simple arrangement for each model, these distributed architectures are considerably different, even though they share such common features as a high-speed network and are often combined into a hybrid arrangements for given business goals and requirements.

Centralized Processing

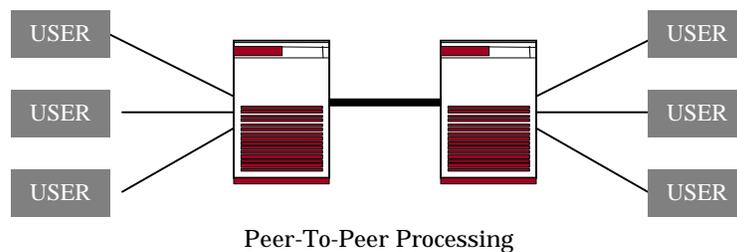
The most commonly used system model in the non-distributed environment is the centralized model, which contains a single server that is responsible for all computing tasks, including user interface and resource management. It is the traditional approach to developing information system and is well understood by designers, developers, and modelers. Users can implement the centralized model on a mainframe, mid-range, or standalone personal computer.



Peer-to-Peer Processing

This was once a very popular and least complex distributed architecture. This architecture extends the centralized concept by adding a second computer that communicates with the first on an equal basis. Each processing computer manages its own data and resources, and the resources on one computer may be shared with the peer computer over a high-speed network. Users are connected to one computer or the other.

The location of resources in a peer-to-peer design is not usually transparent. The local computer, which is the computer where the user interfaces with the application, does not know the location of the other resources. The application must contain this information and make a request to a specific computer for access. If a specific resource is moved to a different computer, it may lead to changes in the local application.

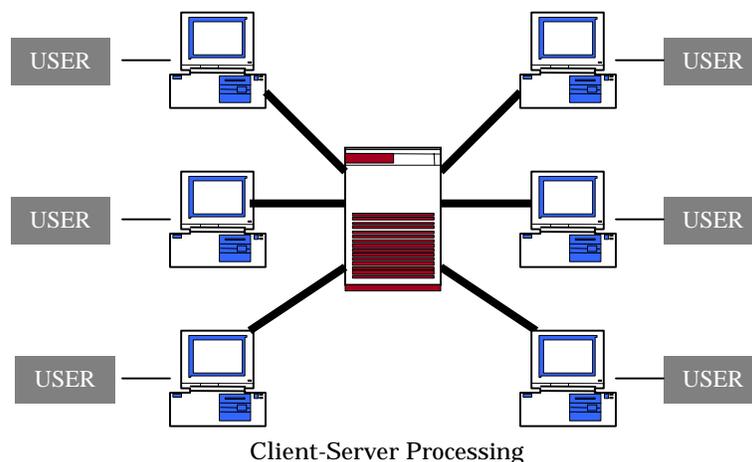


Client/Server Processing

The client/server model has two types of processing platforms, --the client and the server. This model sets up a distinct employer-employee relationship to manage resources and data. The server acts as a master for several clients and, although resources are shared among clients, all requests for sharing are managed by the server.

The server does not, however, do all the processing. It simply controls access to the data from a single point. The client computer, on the other hand, allows for user interface, provides associated processing, and is responsible for private, non- shared data. A developer, manager, and data entry clerk may deal with the same data in entirely different ways. The client computer provides an intelligent user interface that can be custom-tailored to each type of user and, in many applications, performs the majority of the processing.

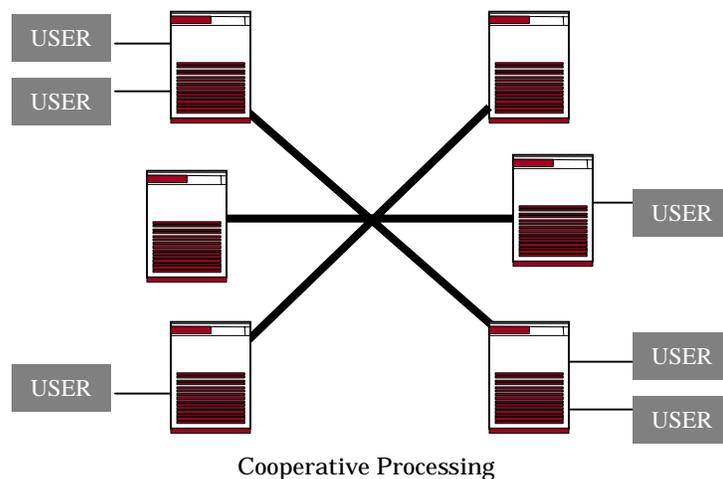
Centralizing resource management functions, including data access, simplifies the design of most data-intensive applications. Concurrency and integrity issues can be minimized. This is one of the reasons that the client/server model was until the advent of the N-tier architecture, the architecture of choice for most data base applications.



Cooperative Processing

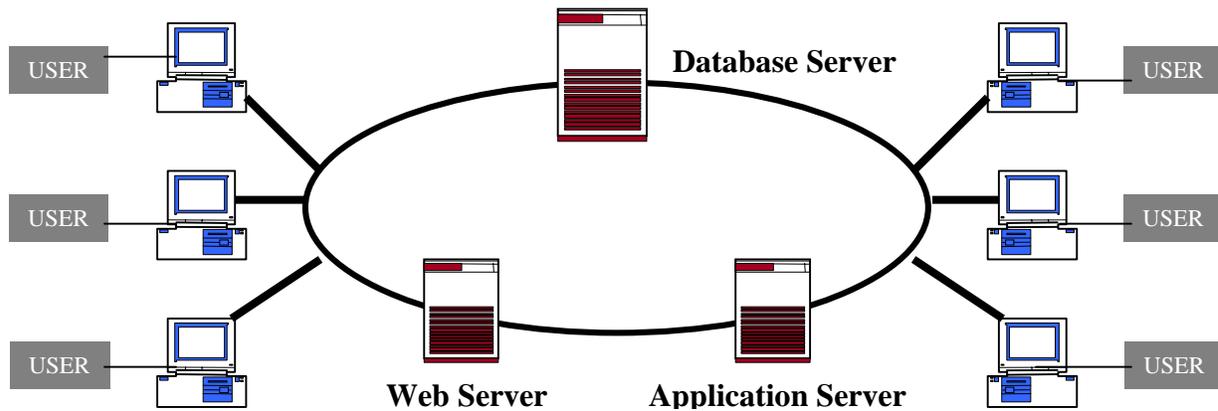
In a cooperative processing network, all computers are connected on a peer-to-peer basis, but the connection is more sophisticated. This allows the resource management function to be a shared task. For example, if database records need to be locked or an update needs to be validated against existing locks, any computer on the network can perform these functions. All processing tasks can also be shared, including the management function. A program on one computer may make a processing request that can be completed by another computer. This allows specialized computers to perform unique processing. This *pool* of computers can respond to processing requests from dozens of other less-powerful workstations.

The cooperative processing architecture also provides a more efficient use of overall computing resources. Idle processors help to balance peak processing loads. The challenge is to accomplish this shared processing in a heterogeneous environment where both the physical hardware and the operating systems differ. Many hardware vendors and standards organizations are working toward this goal, and have raised some rather impressive architectures.



Three to 'N' Tier Processing

The three-tier software architecture emerged to overcome the limitations of the two-tier client-server architecture. The third tier (middle tier server) is between the user interface (client) and the database server components. This middle tier provides process management where business logic and rules are executed and can accommodate hundreds of users by providing functions such as queuing, application execution, and database staging. The three tier architecture is used when an effective distributed client/server design is needed that provides (when compared to the two tier) increased *performance, flexibility, maintainability, reusability and scalability* while hiding the complexity of distributed processing from the user. N-tiers can be added to this model by involving addition servers for the purposes of managing middleware, applications or database access. The most common example being thin client web based applications incorporating web servers that interact with applications and databases.



Distributed Database Architectures

The detailed design and installation of a distributed database architecture deals with a wide range of issues, some of which are covered below. Since so many people will ultimately benefit from or be affected by the distributed database architecture, the selection of the most appropriate hardware and software for the specific needs of the organization is paramount. In addition, once the new system is in place, continuous support is needed to ensure that any problems are handled quickly and that new needs are addressed as they arise.

Three important features characterize the distributed database architecture. These are location and replication transparency, programming language transparency, and multi-site update capability. The availability of these features depends on the software architecture and the capability of the products and tools chosen to be part of it.

Location and Replication Transparency

This feature supports data independence, which enables the database manager to change the physical data structures and access paths used in a database without modifying existing application programs. Distributed database architecture supports location transparency if the user is not aware of the location or site of the data being accessed. Replication transparency is supported if the user is not aware that more than one copy of the data exists. Although this feature provides increased data independence, they require the use of sophisticated and expensive software optimizers.

Programming Language Transparency

Programming language transparency is another important feature. When a distributed database architecture supports language transparency, the user formulates request using a single data manipulation language. A translator in the application package, system hardware or database management system translates the request to the language understood by the databases throughout the architecture.

Multi-site Update Capability

The third important feature is the capability for updating a number of local, regional and central databases in a single request. This requires a database product with sophisticated distributed concurrency control mechanism that guarantee that two or more users do not attempt to update the same data simultaneously. They also require distributed commit

protocols that enable distributed databases to determine if and when updates to the database are completed. The distributed concurrency control and distributed commitment add to the complexity of the distributed database paradigm and to the communications costs and response time of distributed requests.

Database Management Systems

For most commercially available distributed DBMS's, the software consists of a family of products that are available on a variety of hardware platforms. A typical family of products might include:

The basic DBMS and its active data dictionary;

The communications software that is coupled with the DBMS. This software may be available with various levels of capability. For example, the minimal capability would be a protocol for remote data access. The next level of capability would be a gateway for remotely accessing foreign databases or files. (Foreign databases are databases established by other brands of DBMS software.) The truly distributed functional capability would be a communications software product that supports location transparency data accesses and concurrence control. It would also include such features as a two-phase commit protocol for ensuring data consistency.

Some distributed DBMS vendors also offer additional software utilities (e.g., fourth-generation languages, query-by-forms or query-by-examples, fancy report writers, OLAP/Multi-dimensional database tools and data base administration tools for monitoring activities).

Interoperability & Heterogeneous Hardware Support

Large organizations with a history of using information technology have a legacy mix of numerous hardware components provided by various manufacturers to fulfill its operational responsibilities. In this case, the USDA is no exception. When approaching the implementation of the Distributed Database Architecture, an important consideration is the ability of the application databases be flexible enough to be deployed on a number of hardware platforms without addition programming or loss of functionality. As described earlier in this document, this is the very definition of an Open System.

During its lifetime, the USDA is likely to require additional yet-to-be identified technologies to be integrated into its architecture. It is imperative that the USDA is able to easily adopt these new technologies and integrate them into their architecture when the need arises without having to re-architect existing systems, applications, or data-sets. To support this goal, the distributed database architecture must ensure that as many architectural artifacts as possible conform to open standards and are as interoperable as possible.

Scalable VLDB Support

Implementation of distributed database architectures suggests large volumes of data. Estimates of the data volume to be managed by the SCI and Modernization Initiatives are

as high as and could a Terabyte. To efficiently manage this large amount of data in a timely manner and to make full use of available computer resources, multi-processor symmetrical multi-processing (SMP) platforms and massively parallel processing (MPP) capabilities (parallel query, index, load) are essential in the hardware platform and with software able to make best use of these state of the art hardware architectures.

The DBMS and server hardware should also demonstrate scalability, meaning that additional computer resources (CPU and memory) yield corresponding increases in DBMS performance. Without a scalable DBMS with parallel processing capabilities, management of large data stores is highly impractical if not impossible.

High Availability

Uninterrupted service is another essential capability of the distributed database architecture. USDA databases currently operate on reliable hardware, and must continue to do so in the proposed architectures. This high level of database availability should not be allowed to diminish when making use of open systems hardware and database software especially with the introduction of the public to USDA information via the Internet.

It is essential that the distributed database support high availability computing where multiple computers share a set of common disks. This enables a single database to be accessed simultaneously by multiple computers, and the failure of a single computer does not prevent database access from other computers in the hardware cluster.

The DBMS should provide on-line database backup and recovery for 24x7 operations in a manner that does not impair simultaneous database access by users and batch processes. Without support of loosely coupled systems, failure of a single hardware component can compromise system availability for undefined and unacceptable periods.

Support for Large Users Populations

USDA currently employs thousands professionals to fulfill its operational responsibilities. Although it is extremely unlikely that all of these people will connect to a single computer simultaneously, the USDA computer system must demonstrate the ability to support hundreds or thousands of simultaneous users. For this reason, it is essential that the DBMS and server hardware scale to meet the demands of the Web enabled USDA community.

Replication and Remote Computing Facilities

Closely linked with distributed database functionality is the ability to replicate information reliably between remote databases in an easy to manage, predictable fashion. Instead of labor-intensive manual methods, requiring many hours and lines of code, the distributed database must have the ability to automatically replicate data from a single database site to multiple regional databases. Conversely, the architecture must be able to automatically replicate information from multiple regional databases to a single database site for centralized analysis or reporting.

It is essential that the DBMS provide replication facilities to provide automated data replication protected by a two phase commit protocol or similar mechanism to ensure reliable transaction execution or recovery. As many users will work in a "disconnected" mode, the replication facilities should provide support for mobile users and devices in both a web environment and PDAs.

Enterprise Systems Management

The technological footprint of distributed database architectures tends to be large and spread across a wide geography. is likely to support a large environment of computers that may be centrally located or distributed across USDA regional offices. In any event, maintaining and managing these machines presents a serious challenge. For this reason, it is imperative that the USDA DBMS provide remote database administration and SNMP support (simple network management protocol) to enable monitoring and management of distributed databases from one or more administration consoles.

This is a critical feature of the distributed database architecture as given a footprint as large as the USDA service center structure, it would be necessary to increase the number of technical personnel multi-fold to support operation of the architecture. Without DBMS remote database administration tools and support of third party system administration tools, managing a enterprise of distributed databases becomes impractical due to the sheer number of technicians required.

Security

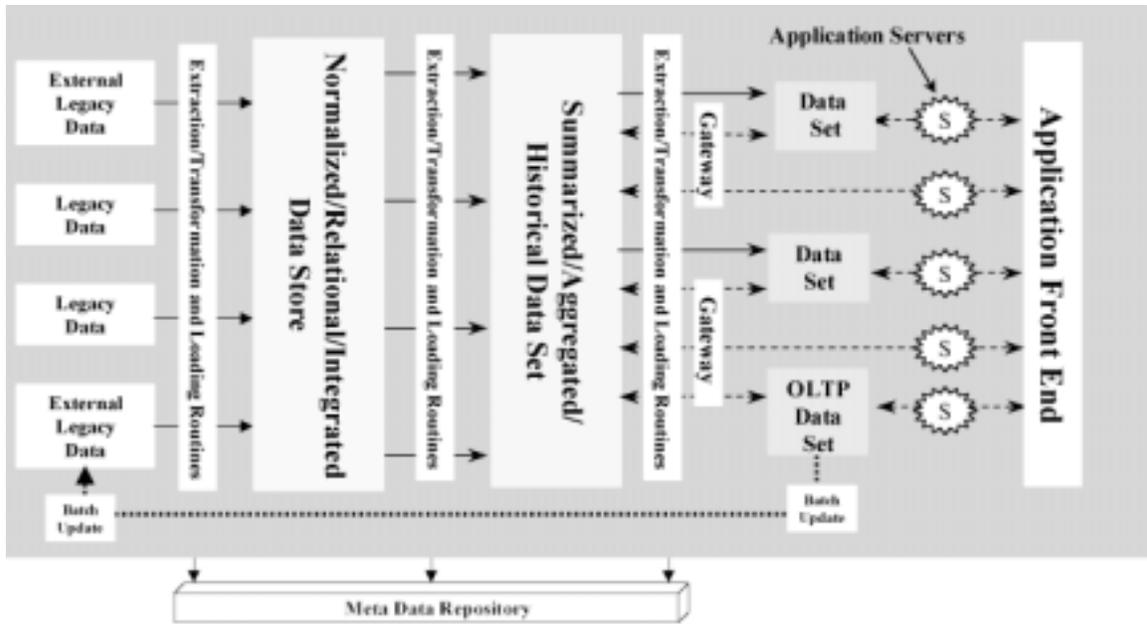
USDA maintains mission-critical systems that store sensitive information about customers, government assets and privileged financial information. As the USDA consolidates service centers and data processing centers, it is imperative that this sensitive information be stored in secure databases that provide data access controls at the table, column, row and operating system level. Additionally, with the introduction of Web technology as the means to provide low cost, universal access to corporate assets, security issues become varied and complex. The proposed tool set for the distributed database architecture must provide rigorous security controls and facilities in accordance with USDA guidelines.

The USDA SCI Data Architecture

With the discussion and main ideas of the *Open, Heterogeneous Distributed Architecture* and the *Distributed Database Architecture* behind us, its time to proceed to the proposed architectural models for the USDA's Service Center Modernization Plan, the Service Center Initiative, and the Geo-Spatial Data Warehouse.

The Conceptual Data Architecture

The clearest way to introduce the following architectural options is to present the premise on which they are all based. This will attempt to be accomplished with the use of the Conceptual Data Architecture depicted below, which is the high level view of the architecture and depicts its principle components (a.k.a. 'artifacts') and the movement of data around and through it. A thorough grasp of this Conceptual Architecture is essential for understanding how the physical architectural models *work*.



The Conceptual Data Architecture

Components of the Conceptual Data Architecture

The Legacy Environment Component

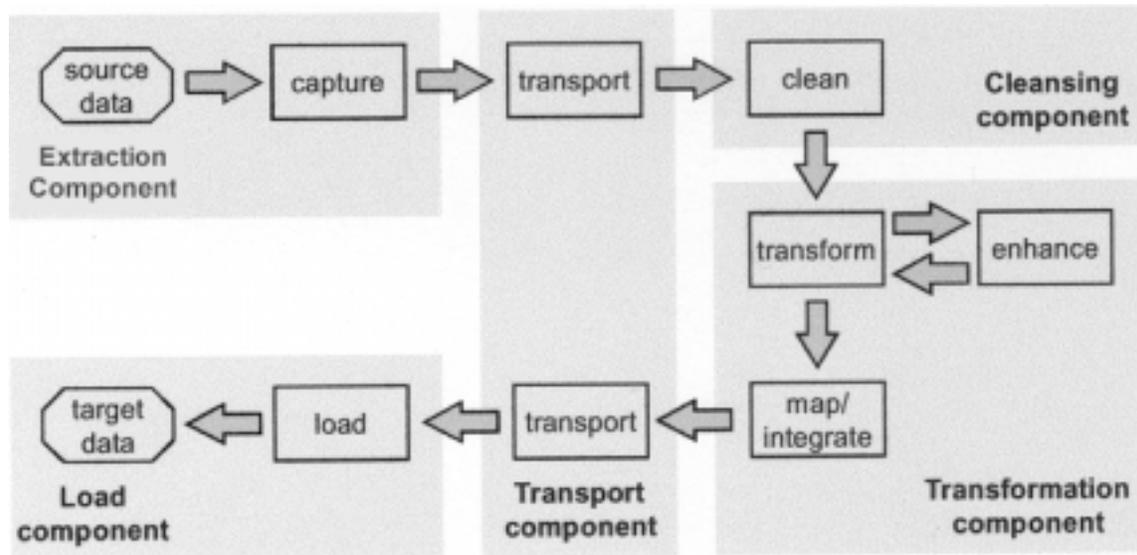
A *legacy data sets and systems* are the existing information processing system that is essential to accomplishing the organization's mission and has the following characteristics:

- It normally represents many years of accumulated experience and knowledge about business operations and problem solving within the organization's business environment.
- It is generally designed as a centralized architecture that exists in various states old to newer hardware and operating system platforms
- It is usually a 'transactional' environment supports information sharing and integrated business processes only within one component of the organization.
- Its software may be the only place where an organization's specialized business rules exist.

The Extraction, Transformation and Loading Process Component (ETL)

One of the key objectives of a data architecture is to put business data in a form that the business user can easily understand and use. The Extraction, Transformation and Loading process does this by extracting or capturing data from source systems, cleaning it, transforming it (e.g. cleansing, enhancing, restructuring and summarizing) and putting it into a form for use (loading) in a target data construct. Accomplishing this objective is by far the most taxing and difficult task(s) in the building of the data architecture.

The level of complexity (number of data sources, amount of data cleanup and transformation) to be dealt with by the ETL process depends on the type of data system being affected (i.e. operational systems, ODS, data warehouse, and/or data marts).



The Extraction, Transformation and Loading Process

There are many off-the-shelf products that support this process, all of which can be categorized as either Code Generators that create tailored ETL programs, Data Replication Tools used for acquiring data from relational database source files and to propagate data changes from a central database to remote ones, or Database Middleware products that provide user access to operational databases.

In addition to integrating disparate data, the ETL process addresses the data quality requirement, which is one of the cornerstones of success in the distributed database environment. In the Conceptual Architecture depicted above, there are four places where the quality of data can be improved as data is moved between the principle data sets in the architecture. These include the legacy environment itself; the operational data store; the historical data set or data warehouse and n the data marts.

The Normalized, Integrated Data Store Component (ODS)

The *Operational Data Store* (ODS) is a tactically structure data set, designed to hold and is generated from production data sources. Unlike its downstream companions, it is a transaction oriented data set with little or no summarization of its contents, hence the term '*normalized.*' The emphasis of the ODS is on access to current or near-current data that is integrated from multiple heterogeneous sources to enable easy access for operational type processes and decisions. Since the ODS generally does not contain historical information, it cannot be used for complex data analysis. Its role is to facilitate the integration of disparate operational systems and upon maturity, and is often positioned as the '*heir-apparent*' to soon-to-be decommissioned legacy data sets.

The Historical Data Set Component (Data Warehouse)

The data warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data that supports the decision making process for an organization. It is literally a combination of subject areas, data sources, user communities, business rules to be applied and architecture. The data in the warehouse is integrated, clean, consistent and reconcilable with operational and legacy data stores.

There is generally a minimum of four levels of data warehouse data, i.e. old detail, current detail, lightly summarized data, and highly summarized data. The data warehouse is the portion of the architected data environment that serves as the single integrated source of historical data.

The most important characteristic of the data warehouse is that it is best used as a point of distribution for accurate, summarized, historical data. The data warehouse is a non-volatile structure meaning that updates to the data contained therein is not done via transactional processes of user interfaces, but is rather fed by predefined loading processes at user defined intervals. The data warehouse also serves well as the source of data for the data mart architecture, from which, the data marts are refreshed at user-defined intervals.

Data Warehouse Constructs

There is much confusion that surrounds the term Data Warehouse. This is largely due to the fact that a data warehouse is not one and only one thing. Data warehouses come in many varieties and sizes, but be categorized as one of the following basic types...

1. A Virtual Data Warehouse model, allows end-users on terminals or client workstations direct access to operational databases and files. This approach provides end-users with query and reporting capabilities against current business information via the use of modern reporting tools, but rarely provides the business information required by users for complex data analysis of any kind.
2. A Decentralized Data Warehouse contains informational data that is of value to a specific user or group of users. Departmental warehouses are synonymous with data marts and contain data captured from one or more operational systems. Frequently operational data is denormalized and summarized before it is applied to a data mart. The data mart approach permits decision support processing to be done on local systems improving both performance and availability. It also provides a quick jump-start for organizations embarking on their first data warehouse project. As the number of data marts grows, however, so does data redundancy, and the complexity of managing the environment. The main disadvantage of a data mart approach is that they provide limited flexibility for satisfying new information requirements.
3. A Distributed Data Warehouse contains several data marts combined together into a single distributed data environment through the use of so-called data base middleware *hub servers*. This approach, though not recommended for first time

builders, has the advantage of incrementally raising the data warehouse in environments where the needs of a specific customer community must be urgently met, and the support and funding for an enterprise-scale central warehouse is not present.

4. A Central Data Warehouse contains integrated informational data captured from one or more operational systems or external information providers. This is the most common approach used to build subject-oriented data warehouses designed for complex data analysis and integration of disparate enterprise data sets. A central warehouse offers more flexibility for satisfying new information requirements and is easier to manage than either multiple data marts or a distributed data warehouse. A central warehouse contains the four type of data described above, transformed from operational systems at user defined intervals.
5. A Two-Tier Data Warehouse employs both the central data warehouse model, and the decentralized data marts. This approach combines the advantages of a central data warehouse and decentralized data marts. Organization already possessing the centralized DW evolve to this model as their architecture mature

Properly constructed, any of these data warehouse constructs will provide some measure of consolidation and integration of an enterprise's business data, and create a virtual "one-stop-shopping" data environment.

The Segmented Data Set Component (Data Mart)

The *Data Mart* contains a subset of enterprise data that is of value to a specific group of users. This subset of data may be captured from one or more data warehouses, operational systems, external information providers or local data sources. The main difference between a data warehouse and a data mart is that the data mart contains a subset of the data warehouse data, and the data tends to have a higher degree of summarization as compared to any other data set in the architecture. It is very important to understand that a data mart is defined by the functional scope of its users, and not the size of the database involved.

The Data Management Component

The Data Management Component consists of a set of services for managing the architecture's data. The services include...

- Authorization Services used for controlling access to warehouse data
- Archive Services for moving data into a near on-line state
- Backup and Recovery Services for handling tehe backup and recovery of data in the event of media failure
- Data Distribution Services for distributing data to other operating environments
- Monitor Services for monitoring and tuning the performance of data movement process within the architecture

The principal construct in the *Data Management Component* is the *Meta Data Repository* and the *MetaModel* it supports. The USDA Data Management team has put together a strong Metadata Repository which is sufficiently extensible to support the requirements of this Data Architecture. In the interests of brevity, this document will defer to the documentation supplied by the Data Management organization for describing the benefits and workings of the Meta Data Repository.

The Data Access Component (Application Front-Ends, Servers and Gateways)

The remaining component of conceptual architecture is the data access component or application layer, which provides the database middleware, gateways, and *front-end* tools that enable users to access and analyze the architecture's data sets.

There are many types of database middle used to access data mart and warehouse database servers from end-user client workstations. These include a) point- to-point middleware servers that allow workstation users to directly access a central data warehouses and marts; b) workstation based hub servers that employ robust metadata schemes that allow data to be access across multiple database servers.

Data access tools range from query generation tools to multi-dimensional products for complex data analysis, and data mining tools that allow users to drill down through mart/warehouse data for information discovery applications.

How the Conceptual Architecture Works

In summary, the conceptual data architecture depicts the movement of data from legacy environments, through a series of extraction, transformation and loading routines, into target data sets on which data access constructs facilitate access and use of the business' data assets.

The data sets that exist at various levels in the architecture can be constructed in various ways depending on the needs and resources of the business. The architecture is flexible enough to tolerate iterations in its construction, yet rigid enough to limit the number of things to be considered during the decision making process.

Critical Success Factors

- Adequate performance of the overall distributed process infrastructure and of the distributed database applications and tool set
- Greater user satisfaction with newly installed applications than with the previous non-distributed, disparate data architecture and applications
- Lower cost for resources and better efficiency in the processing of
- Improved maintenance of the distributed data base system and the distributed data administration/management processes.

Introduction to the Architectural Alternatives

Challenges of Enterprise-wide Systems Architecture

Under the best of circumstances, the definition of an enterprise-wide architecture is a daunting task. The raising of new applications and the integration of disparate data sets to reduce the total cost of computing to the organization and improve overall efficiencies, challenges even the most experience architects and designers. In addition to leveraging everything in the current infrastructure—hardware, applications, networks and talents, other considerations include...

- Managing and supporting users in a timely and cost-effective manner
- Extending access to business-critical applications to dispersed user—regardless of connection, location or device
- Ensuring acceptable performance of the overall architecture
- Providing security in the necessary measures

USDA Service Center Initiative Technical Architecture Assumptions

The USDA has been gathering user requirements for a considerable period. Based upon a general understanding of those requirements the following set of architectural assumptions have been gleaned...

- It will be Network-centered and built on protocols and technologies that allow for some level of seamless internet access and information exchange.
- The hardware will support high data volumes.
- Remote access and mobile computing requirements must be accounted for.
- The environment is heterogeneous composed of the Microsoft's Windows operating systems, large UNIX database servers, mainframes, and a variety of third party applications and tools.
- Based on distributed transaction models and ad hoc data production.
- The architecture is component-based and executed in a distributed software and hardware environment
- The database architecture must be scaleable and run on a variety of platforms
- The architecture must be manageable from central locations for efficiency, and operate at mainframe levels of stability.
- The architecture is open and supportive of interoperability

Based on the requirements for information access and exchange, it can be assumed that mail enabled applications and message brokers will provide services that can coordinate and route messages between geographically dispersed clients.

USDA Service Center Initiative's Technical Constraints

Assessing the technical infrastructure of the USDA's Service Center Initiative reveals a number of constraints that unless rethought and remedied, will force the implementation of an architecture that very nearly mimics the existing legacy environment. This could require the abandonment of many of the stated system requirements and result in the loss of many of the benefits inherent in the distributed architecture paradigm.

The constraints found here are not listed in any particular order, but those of most significant severity will appear towards the top of the following list...

1. Network Bandwidth: The network bandwidth currently stands at 56kbps for the majority of the Service Centers in the SCI architectural footprint. Some sites have not yet been upgraded to this speed and remain at 28.8kbps, using dial-up modems. It is estimated that the 58kbps wire, though only recently installed, is already nearing saturation. Essential to Distributed and web based architectures is high speed network. This lack of bandwidth represents the most significant challenge to the goals of the Service Center Initiative.
2. Network Configuration is Unknown: It is not certain how the current installed network is configured.
3. Big Data and Small Processors: The pilot environment is intended to prove the concept of distributed architecture, demonstrating the capability to communicate and move data across large geographies. The platforms chosen to-date tend to be relatively small, Intel based architectures, driving small scale distributed databases. Under the best of network conditions, large processors with large back-planes are required to support large user communities using several application packages. CCE continues to model larger scale platforms for use in a national technical infrastructure.
4. Unsettled Software Decisions: Some of the main software components, namely the Data Base Management System, Middleware software, ETL software and archival and retrieval software.
5. Absence of Analytical Constructs: The constructs mentioned earlier in the document (i.e. the enterprise data model, data usage models, data distribution models, volumetrics, denormalization maps, etc...) are extremely modest in their composition and in most cases do not exist at all.

Building Blocks for the Architectural Alternatives

The alternatives presented here are based on three principle ideas. The first is the principle of server cooperatives. It is the guiding principle inherent in each of the recommended variations. The second is the Server-Based computing paradigm as defined by Citrix, which addresses the lack of a high speed network in the USDA infrastructure. Lastly, is the Information Hub and Spoke construct which directly addresses the distribution of data across the enterprise footprint.

Server Cooperatives

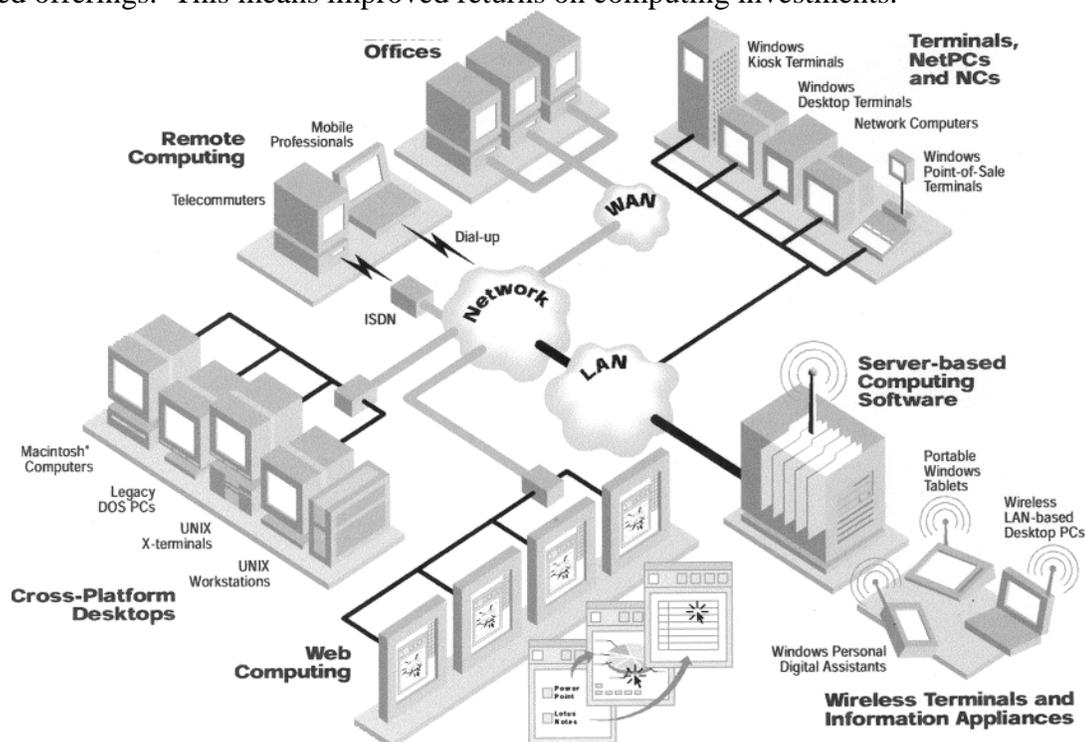
Server Cooperatives is a principle based on server consolidation, which enables the simplification of the server architecture by rolling in big boxes to replace armies of little ones, and moving systems to data centers under the care of skilled IT baby-sitters. Beyond improved management and better service, server consolidation can mitigate increasing IT headcount, save money in the management tools budget, and allow IT to shift resources from operations to development.

In addition, as new application technologies continue to gain prominence, (especially internet and e-commerce apps) the architectures used to move information throughout the enterprise must be able to scale up to accommodate new features and functionality. Server Cooperatives or server consolidation will allow the IT organization to use technologies that not only scale but also improve overall service to the business and the business partners. Such technologies include the entire range of operating systems and their packaged/layered products; Symmetrical Multiprocessing and Massively Parallel Processing technologies; shared disk subsystems which ease the storage management task, and high-speed communication inter-connects between servers and the outside world.

Server Based Computing

Server based computing is a model in which applications are deployed, managed, supported and executed 100% on a server. It uses a multi-user operating system and a method for distributing the presentation of an application's interface to a client device.

Within the server-based computing paradigm, client devices, whether "fat" or "thin," have (relatively, based on network infrastructure) instant access to business-critical data and applications via the server—without the need for application rewrites or downloads. In addition, server-based computing works within the current computing infrastructure and current computing standards, and with the current and future family of Windows-based offerings. This means improved returns on computing investments.



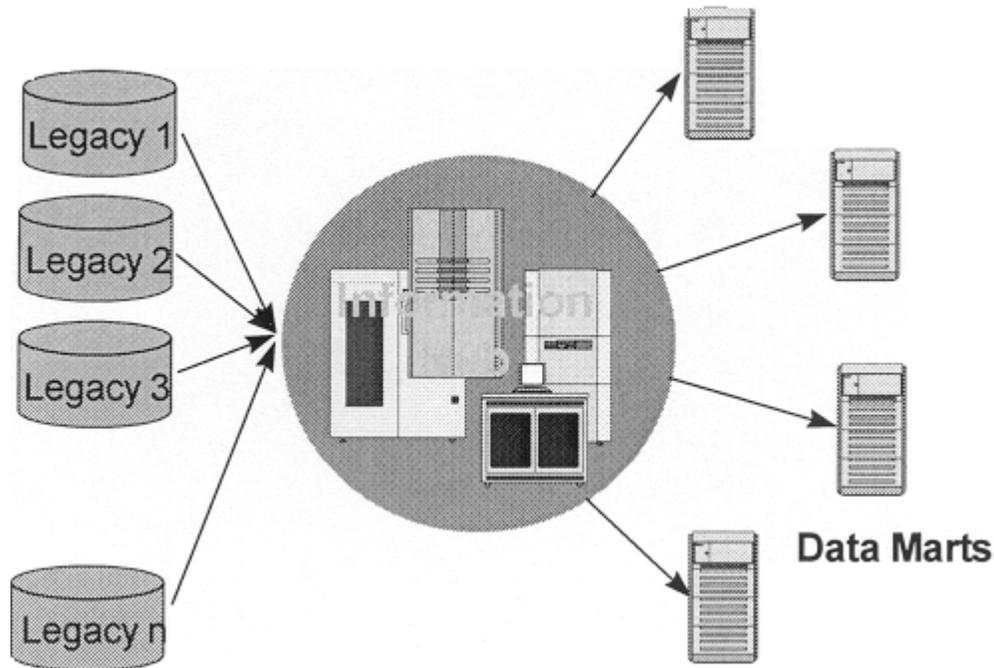
In short, the main benefit to the USDA of this paradigm is the leveraging of a relatively slow network infrastructure. The server-based computing paradigm is the most reliable way to reduce the complexity and total costs associated with the goals of enterprise computing^v.

The Information HUB and Spoke Construct

The Hub's role is to collect, manage and disseminate data to those users and applications that require it. The HUB contains atomic-level data, summarized level data, and any data shared by more than one application or user group live on the Hub.

The Hub is a place to manage and store all the data an enterprise shares. The Hub and Spoke architectures moves data throughout the enterprise – from those service points that collect it, to end user systems that turn the data into information.

The Hub and Spoke architecture is designed to centralize systems management and operational processes while supporting the information distribution requirements of a distributed computing environment. The Hub enables large architectures to overcome limitation inherent in conventional warehouse technologies, primarily in the areas of cost and scalability. Other benefits include reduction in extraction complexity; reduction in data redundancy, leveraging of existing computer environmental resources and the establishment of common transformation rules and processes.



The Hub and Spoke Construct

Major Data Constructs in the Architectures

In the models that follow, the principle data constructs (e.g. the Operational Data Store (ODS), the Data Warehouse (DW) and the Data Marts (DM). *Please refer to the Conceptual Model*) are the focus of difference between the models. As a rule, the ODS will always exist nearest the legacy data set, while the data warehouses and data marts will be distributed in different configurations. *Note:* The data mart should always be assumed to be the driving data set for the Server-Based computing (SBC) construct

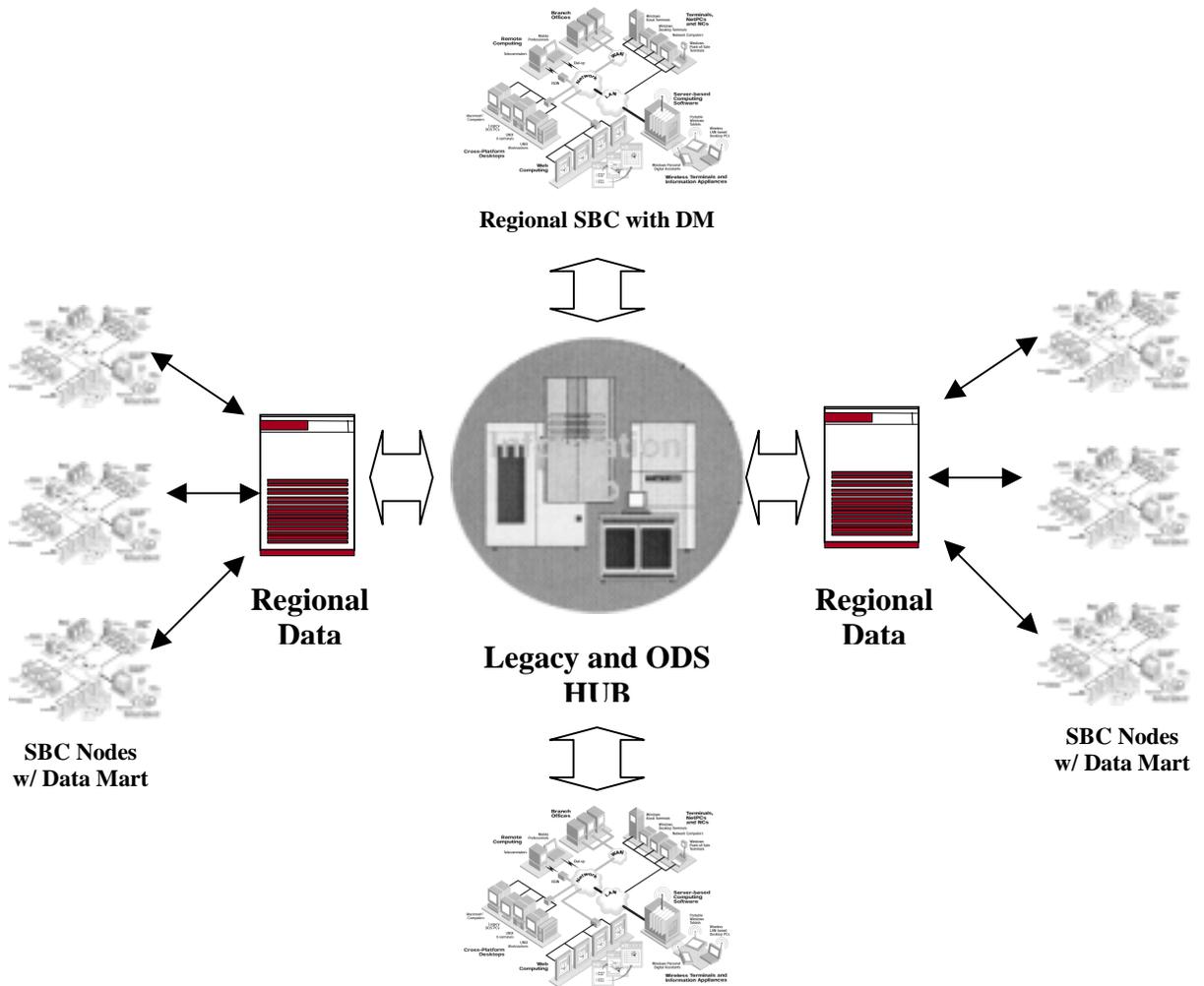
Data Movement in the Architectures

Using the Conceptual Model as a guide, data movement is always from legacy towards the data marts, going through a series of ETL transitions, with return loops to the ODS and/or the legacy environment from transactional data mart and applications as required.

Other Major Assumptions in the Architectures

1. The super-set of shared data lives exclusively at the ODS level of the architecture, with the exception of the enterprise-wide data warehouse scenario where the data warehouse is a shared, enterprise resource.
2. Wherever multiple data warehouses appear, the data in those warehouses is a sub-‘horizontal partition’ or sub-set of data, consisting of data for a specific region and/or USDA agency.
3. The data warehouse is *never* a transactional data store. It receives its updates directly and exclusively from the operational data store, legacy data stores, or external data sets via batch mode or processes.
4. Data Marts can be either transactional or query only.
5. The architectures are a hybrid of centralized, cooperative, and multi-tier configurations.

Configuration Three: Hybrid with Call Center Architecture



Appendix A: The Data Management Architecture

According to the *Data Administration Concept of Operations*,^{vi} data architecture is “an orderly arrangement of Service Center data resources to achieve a:

- 1) common understanding of data resources available;
- 2) planned approach to data acquisition, storage, and retrieval to achieve a high degree of responsiveness to user demands; and
- 3) high degree of data sharing and data mobility to reduce program delivery costs.”

Defining the location and distribution for enterprise data is the data architecture task that most effects the technical architecture. The location and distribution of enterprise data is dictated by business needs and technical issues. Specific objectives of the data architecture task are to:

- 1) define the physical architecture of the data, i.e., location of data across the network;
- 2) describe data and databases located on a central server, decentralized servers, and local clients;
- 3) maximize the functionality and responsiveness of software systems by making use of metrics to assess movement of data through the USDA network;
- 4) maximize the functionality and responsiveness of software systems by accounting for where and how often data is updated or refreshed; and
- 5) promote a high level of data sharing and data mobility.

Through BPR work already completed, the following categories of data assets, shown in Exhibit 1, have been identified:

- Common data is data jointly owned, used, and managed by Service Center partners. The common data sets of interest for the BPR initiative are customer data, office data, administrative data, land-unit data, and standard geospatial data.
- Shared data is data owned and managed by a specific Service Center partner but shared by other partners. The shared data sets of particular interest for the BPR initiative are the natural resource data sets, specifically soils, plants, climate, and demographic data.
- Unique data is data owned and managed by a specific Service Center partner and not shared.

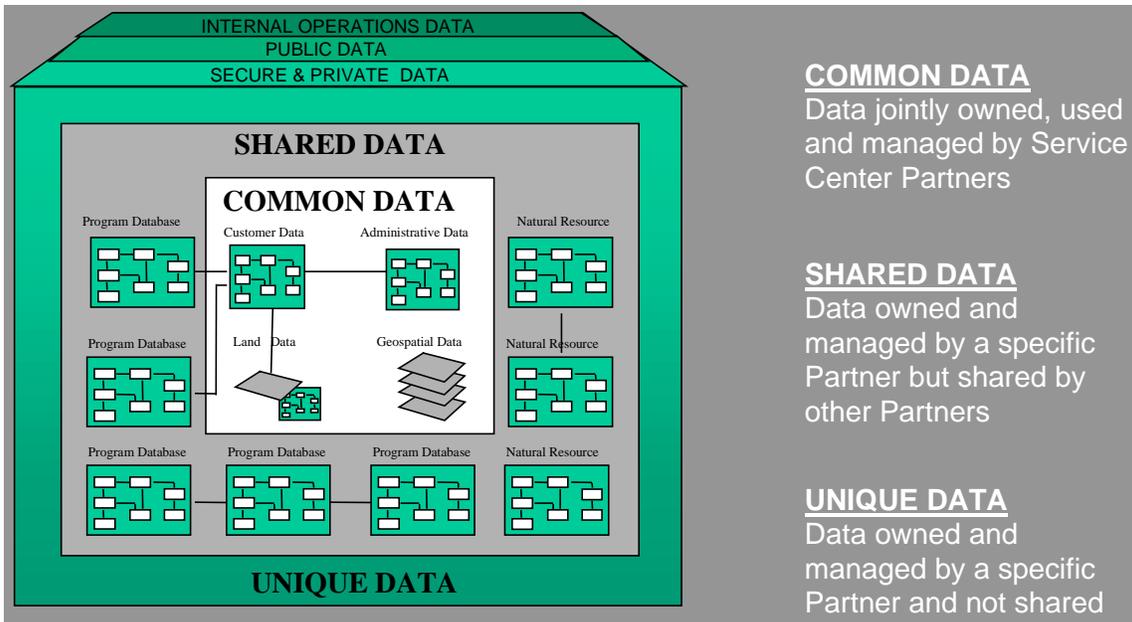


Exhibit 1: Service Center Enterprise Data

In conformance with the n-tier application architecture, the Data team is encouraging the development of reusable software modules (components) to encapsulate access to common data (jointly owned, used and managed). All applications that need access to the data would do so via the provided components. The key advantage of this approach is increased data integrity. As illustrated in Exhibit 2, all changes to data (creates, reads, updates and deletes) are handled by the components, allowing business rules to be implemented and enforced in one place. This also relieves individual application development projects from having to write their own data access routines.

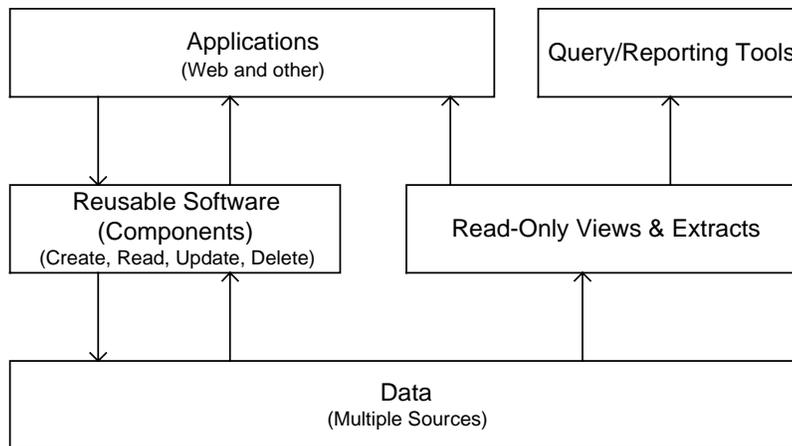


Exhibit 2: Managing Data Access^{vii}

The above descriptions provide a logical view of some of the data architecture issues. The SCI Data Management team is still developing specific data placement and distribution alternatives and strategies that will affect the target

technology architecture. Specific BPR projects, such as the Service Center Information Management System (SCIMS) project, are the first pilot applications to test and provide feedback for the various data placement configurations.

Appendix B: CCE Operating System Software

CCE software covers a broad range of functionality, from the operating system to office automation to enterprise based software like geographic information systems. The following sections detail the current status and selection of each of the researched CCE software product categories.

Workstation Operating System (O/S)

Selection Status Summary: Windows NT Workstation 4.0 software was selected for use on laptops and desktop workstations in the BPR pilot sites and the Y2K replacement initiative.

CCE Product Category	BPR Pilot Sites Configuration	Phase 1 (FY98 and FY99)
Laptop O/S	Windows NT Workstation 4.0	Windows NT Workstation 4.0
High-End Desktop O/S	Windows NT Workstation 4.0	Not Applicable (none purchased)
Mid-range Desktop O/S	Windows NT Workstation 4.0	Windows NT Workstation 4.0

Exhibit 3: BPR and Y2K Operating Systems

Reference: CCE Pilot Sites and Y2K Configurations, dated 27 October 1998.

Network Server Operating System

Selection Status Summary: Windows NT Server 4.0 software was selected for use in the BPR pilot sites.

CCE Product Category	BPR Sites Configuration
Network Server O/S	Windows NT Server 4.0 (10 User License)

Exhibit 4: BPR Pilot Site Network Operating Systems

References: CCE Pilot Sites and Y2K Configurations, dated 27 October 1998 and LTD Results Report, Phase II, dated 9 April 1998.

Application/Database/GIS Server Operating System Platforms

Selection Status Summary: As software products are selected and additional technical architectural analyses are conducted, decisions will be made regarding these server choices. Depending on the application and data distribution architectures these servers may reside on one or more physical machines.

Application Server

Selection Status Summary: **Windows NT Server** has been initially selected for Service Center level servers. However, no specific application server platform has been selected for servers running applications outside of the Service Center. Refer to Section 4.2.3 (Application/Database/GIS Server Hardware) for more details.

Database Server

Selection Status Summary: No Operating System selection has been made. This selection is dependent on DBMS selection and application architecture.

Geographic Information System (GIS) Server

Selection Status Summary: No Operating System selection has been made but the candidates are Windows NT and UNIX. The GIS server must be compatible with the ESRI Enterprise GIS product suite.

Enterprise Management System (EMS) Software

Selection Status Summary: No final product has been selected. Final Selection of EMS Products will be made after further analysis of the products, USDA needs, and available funding. The CCE Team is investigating the use of **Microsoft System Management Server (SMS)** to provide initial functionality while other options are being investigated.

Test and Evaluation Summary

As a prerequisite step, business and technical requirements for EMS software were developed and reviewed by USDA Partner Agency contacts, Interoperability Lab staff, and Common Computing Environment (CCE) staff. The requirements fall under the following major functional/technical categories:

- Software Distribution Capabilities
- System Hardware/Software Inventory Functionality
- Problem/Fault/Operations Management
- Performance Management
- Security Management

- Configuration Management
- Distributed Database Management
- Help Desk Support
- Backup and Recovery Management
- Windows NT Specific Requirements
- General Requirements

A market survey of leading enterprise management software (EMS) vendors was performed to determine whether each vendor product met the individual defined requirements. The survey evaluated the business viability of vendors, as well as high level functional and technical capabilities of their software products. The candidate EMS products that were evaluated included:

- Cabletron System's *Spectrum Enterprise Manager*,
- Computer Associates' *Unicenter The Next Generation (TNG)*,
- Hewlett Packard's *OpenView Enterprise Edition*,
- Microsoft's *Systems Management Server*, and
- Tivoli System's *Enterprise Manager*.

Two primary approaches for EMS selection/implementation were utilized:

- a) Use of *best of breed point products*: Point products concentrate on a particular niche EMS function(s), sometimes providing the "best" market product for that function.
- b) Use of a *single vendor comprehensive product*. The single vendor option attempts to fulfill as many EMS requirement areas as possible via a single vendor, thus reducing integration issues.

Initially USDA decided to utilize a comprehensive "enterprise" level package, with a single vendor being the primary provider for EMS software. This places the burden of integrating varying EMS functionality along with accountability for all EMS functionality on the vendor as opposed to USDA.

Another primary function reviewed was the EMS product's ability to support technology components currently being investigated by the Common Computing Environment team. For example, USDA requirements stipulate that application server platforms (including USDA legacy server platforms), and multiple database management system products (e.g., IBM DB2, Informix, Microsoft SQL Server, Oracle, Sybase) be supported.

Based on the requirements analysis and market survey, it was determined that there are two vendors currently capable of delivering an EMS software package that can meet USDA Enterprise Management Software requirements: Computer Associates and Tivoli Systems. The other three vendors (Cabletron, Hewlett Packard, and Microsoft) support a majority, but not all of the requirements.

Because EMS encompasses multiple categories, CCE is in the process of determining whether a single vendor approach or the best of breed approach will be most cost effective in the long term. This technology category will be researched further once other CCE components (e.g., DBMS) are officially selected and after CCE evaluates EMS in comparison to alternatives such as seat management. In the interim, the CCE team is investigating the use of Microsoft Systems Management Server (SMS), a part of BackOffice, as a potential initial or interim solution while other options are being investigated. Details of this evaluation may be found in *CCE Analysis of MS SMS as an Enterprise Mgt. System, Draft, December 1999*.

References: Enterprise Management Software Market Survey, Draft, dated 17 May 1999. Pp.1-3.

CCE Analysis and Testing Update, dated 3 June 1999. Pp. 3-4.

CCE Analysis of MS SMS as an Enterprise Mgt. System, Draft, December 1999.

CCE Hardware Configurations

Decisions regarding standard Common Computing Environment (CCE) hardware/software configurations have been made for the Okeechobee/Gainesville pilots, eight additional BPR pilot sites and CCE Phase 1 deployment. CCE Phase 1 included purchases in both FY98 and FY99. The CCE Phase 1 purchase in FY98 included Y2K replacements as well as initial CCE workstations. The CCE Phase 1 for FY99 provided additional desktops, laptops, and printers. The standard BPR pilot configurations have been selected for network server, high-end desktop, mid-range desktop, laptop and accessories, office automation software and several types of printers (i.e., portable color printer, postscript laser monochrome printer and large format color printer). The Phase 1 configurations included mid-range desktops, laptops, accessories, and office automation software.

Workstation Hardware

Laptop

Selection Status Summary: Laptop hardware was selected and purchased in 1999 for use in BPR pilot sites and the CCE Phase I deployment.

As shown in Exhibit 5 and Exhibit 6, laptop configurations have been selected for BPR pilot sites; their configurations are presented in the first and second columns. In addition, laptops have also been selected in the CCE Phase 1 FY98 and FY99.

Okeechobee/Gainesville Pilot Configuration	Additional BPR Sites Configuration	CCE Phase 1 (FY98)
Dell Latitude CP M233XT 233 Mhz Pentium 64 MB RAM	Dell Latitude Cpi266XT 266 Mhz Pentium II 128 MB RAM	Dell Latitude Cpi266XT 266 Mhz Pentium II 64 MB RAM

12" Active Matrix LCD 20X CD-ROM 2 GB Hard Disk Drive 33.6 KBS Modem Carrying Case Docking Station with Integrated Ethernet NIC Ports: Serial, Parallel, Video, Keyboard, 17 " monitor Mouse Security Lock	13.3" Active Matrix LCD 20X CD-ROM 4 GB Hard Disk Drive 33.6 KBS Modem Carrying Case Docking Station with Integrated Ethernet NIC Ports: Serial, Parallel, Video, Keyboard, 17 " " monitor Mouse Security Lock	13.3" XGA Active Matrix Display touch pad 24X CD-ROM 3.2GB Hard Disk Drive 56 K PCMCIA Modem C/Dock Expansion Station with Integrated Ethernet NIC 17" Monitor Keyboard Mouse 3 Year On-Site Parts & Labor Warranty
--	--	--

Exhibit 5: Pilot Site/Phase 1 (FY98) - Laptop Configurations

Phase 1 (FY99) with Port Replicator	Phase 1 (FY99) without Port Replicator
Dell Latitude CpiR400GT400 MHz Pentium II 128 MB RAM 4.8 GB HD 24X CD-ROM 14.1" XGA Active Matrix Display Touch Pad 1 Serial Port 1 Parallel Port 17" 1024x768 Dell M770 Monitor PS/2 104 Key Keyboard PS/2 Mouse 3Com V.90 56K XJACK WIN Modem PCMCIA Card 3Com 10/100 PCMCIA V2.0 LAN Card Enhanced Port Replicator with Monitor Stand & Integrated 10/100 3Com LAN Card Nylon Carrying Case MS NT 4.0 Operating system	Dell Latitude CpiR400GT400 MHz Pentium II 128 MB RAM 4.8 GB HD 24X CD-ROM 14.1" XGA Active Matrix Display Touch Pad 1 Serial Port 1 Parallel Port 3Com V.90 56K XJACK WIN Modem PCMCIA Card 3Com 10/100 PCMCIA V2.0 LAN Card Nylon Carrying Case MS NT 4.0 Operating system

Exhibit 6: Pilot Site and CCE Phase 1 (FY98 & FY99) Laptop Configurations

Reference:

CCE Pilot Sites and Y2K Configurations, dated 27 October 1998.

CCE Home Page - <http://www.usda.gov/servicecenter/cce/index.html>

High-End Desktop

Selection Status Summary: High-end desktop hardware was selected and purchased in 1999 for use in BPR pilot sites.

High-end desktop configurations have been selected for the BPR pilot initiatives. Exhibit 7 shows the Okeechobee/Gainesville pilot configuration, the configuration for the additional BPR pilot sites, and CCE Phase 1 (FY99).

Okeechobee/Gainesville Pilot Configuration	Additional BPR Sites Configuration	CCE Phase 1 (FY99)
Dell 266 Mhz Pentium II with Integrated 3COM Ethernet NIC 128 MB RAM 6.4 GB IDE Hard Disk Drive 19" monitor 2X AGP Graphics Controller with 4MB Video Memory Upgrade (for a total of 8MB) 12/24X CD-ROM Sound Card Altec Lansing ACS90 Speakers	Dell 400 Mhz Pentium II with Integrated 3COM Ethernet NIC 256 MB RAM 6.4 GB IDE Hard Disk Drive 19" monitor 2X AGP Graphics Controller with 4MB Video Memory Upgrade (for a total of 8MB) 12/24X CD-ROM Sound Card Altec Lansing ACS90 Speakers	DELL Precision Workstation 610 500 Mhz Pentium III 256 MB RAM/100MHz bus (Dual Processor Upgradable) 18 GB Ultra II SCSI HD 17X/40X CD-ROM 10/20 SCSI TR-5 NT PWS Tape Backup 21" 1600x1200 DELL UltraScan 1600 HS Monitor Diamond Viper 770D 2X AGP Video/ w 32MB SGRAM PS/2 104 Key Keyboard & PS/2 MS IntelliMouse 2 Serial Ports 1 ECP Parallel port 2 USB ports Sound Blaster Pro 16bit Audio Sound Card with Hamon/Kardon 195 Speakers 3Com 3C18 10/100 PCI Ethernet Card MS NT 4.0 Operating system.

Exhibit 7: Pilot Sites High-End Desktop Configurations

Reference:

CCE Pilot Sites and Y2K Configurations, dated 27 October 1998.

CCE Home Page - <http://www.usda.gov/servicecenter/cce/index.html>

Mid-Range Desktop

Selection Status Summary: Mid-range desktop hardware was selected and purchased in 1999 for use in BPR pilot sites and the Y2K replacement initiative.

Mid-range desktop configurations have been selected for the BPR pilot initiatives as well as the Y2K initiative.

Exhibit 8 shows the Okeechobee/Gainesville pilot configuration and the configurations for the other BPR pilot sites, along with the CCE Phase 1 FY98 and FY99 configurations.

Pilot Configuration	CCE Phase 1 (FY98)	CCE Phase 1 (FY99)
----------------------------	---------------------------	---------------------------

<p>Okeechobee/Gainesville Pilot Configuration Dell 266 Mhz Pentium II with Integrated 3COM Ethernet NIC 64 MB RAM 6.4 GB IDE Hard Disk Drive 17" monitor 2X AGP Graphics Controller with 4MB Video Memory 12/24X CD-ROM Sound Card Altec Lansing ACS90 Speakers</p> <p>Additional BPR Sites Configuration Dell 400 Mhz Pentium II with Integrated 3COM Ethernet NIC 128 MB RAM</p> <p>6.4 GB IDE Hard Disk Drive 17" monitor 2X AGP Graphics Controller with 2MB Video Memory Upgrade (for a total of 4MB) 12/24X CD-ROM Sound Card Altec Lansing ACS90 Speakers</p>	<p>Compaq Deskpro EP 400 Mhz Pentium II 10/100 Ethernet NIC 64 MB RAM 6.4 GB Hard Disk Drive 17" monitor AGP Graphics with 4 MB RAM 32X CD-ROM Sound Card Speakers Keyboard Mouse</p>	<p>Gateway E4200-450 450 Mhz Pentium III, 64 MB RAM/100MHz bus 13.6 GB HD 17X/40X CD-ROM 17" 1024x768 DaeWoo 712D Monitor ATI Rage 128GL AGP Video/ w 16MB SGRAM PS/2 104 Key Keyboard & PS/2 MS IntelliMouse 2 Serial Ports 1 ECP Parallel port 2 USB ports Sound Blaster AudioPCI 64D Sound Card with GCS-200 Cambridge Speakers 3Com 10/100 PCI Ethernet Card MS NT 4.0 Operating system</p>
--	--	---

Exhibit 8: Pilot Sites and Phase 1 (FY98 & FY99) Mid-Range Desktop Configurations

Reference:
 CCE Pilot Sites and Y2K Configurations, dated 27 October 1998.
 CCE Home Page - <http://www.usda.gov/servicecenter/cce/index.html>

Network Server Hardware

Selection Status Summary: Network server hardware was selected and purchased in 1999 for use in BPR pilot sites.

Exhibit 9 lists Network Server configurations for the Okeechobee/Gainesville and the additional BPR pilot sites

Okeechobee/Gainesville Configuration	Pilot	Additional BPR Sites Configuration
Dell 266 Mhz Pentium II (single Processor-dual processor capable) 128 MB RAM RAID 5 with 3 9GB SCSI Hard Disk Drives 12/24 GB DDS-3 Tape Backup Ethernet NIC 12/24 CD-ROM 17" monitor Keyboard Mouse 1.44 MB Floppy drive Smart UPS		Dell 333Mhz Pentium II (single processor-dual processor capable) 256 MB RAM RAID 5 with 3 9GB SCSI Hard Disk Drives 12/24 GB DDS-3 Tape Backup Ethernet NIC 12/24 CD-ROM 17" monitor Keyboard Mouse 1.44 MB Floppy drive Smart UPS

Exhibit 9: Network Server Pilot Configurations

Reference: CCE Pilot Sites and Y2K Configurations, dated 27 October 1998.

Application/Database/GIS Server Hardware

Selection Status Summary: As software products are selected and additional technical architectural analyses are conducted, decisions will be made regarding these server choices. Depending on the application and data distribution architectures these servers may reside on one or more physical machines.

Application Server

Selection Status Summary: Windows NT Server on an Intel-based machine was selected for BPR Service Center Pilot Sites. No specific hardware product has yet been officially selected for other sites.

Test and Evaluation Summary

The Application Server decision is closely related to other decisions such as the selection of an Enterprise GIS product; legacy migration strategy; and the enterprise application architecture (e.g., client/server, Web-based, etc.).

The *USDA Service Center Business Need and Technical Alternative Evaluation Study – Phase II, April 9, 1998* identified 19 options for application server configurations. These were later reduced to the following three options:

- Windows NT Server - Local level
- Unix Server – State/Regional Level
- AS/400 – State/Regional Level

Based on a number of factors including the benefit cost analysis completed for all candidate options which evaluated both hardware and legacy conversion costs, recent decisions by FSA to pursue legacy connectivity independently of the CCE effort, and requirements of the GIS architecture to locally store large GIS files, an Intel-based server running Windows NT Server was initially selected as the application server environment for the Service Center level, including BPR pilot sites. However, no specific application server hardware platform has yet been selected for servers running applications outside of the Service Center.

Database Server

Selection Status Summary: No Hardware Product selection has been made. This selection is dependent on the Database Management System (DBMS) selection.

Geographic Information System (GIS) Server

Selection Status Summary: No Hardware Product selection has been made. This selected server hardware, however, must be compatible with the ERSI Enterprise GIS product suite which has been selected for initial deployment. ESRI's server compatible products work with Windows NT and UNIX operating system software. No decision has yet been made regarding the GIS server operating system.

Performance modeling will be conducted to help size the GIS servers.

Reference: CCE Updated Benefit Cost Analysis, dated 11 June 1999.

Mobile Computing

Selection Status Summary: No products have been selected. Candidate products in the sub-notebook category were identified in a draft market survey report.

Test and Evaluation Summary

Requirements collected for mobile computing components initially concentrated on personal digital assistants (PDA). The requirements received pointed out the need for multiple categories for mobile computing:

- Sub-notebook computers – small, lightweight notebook computers for mobile users.
- Hand-held data collection devices – ruggedized hand-help single purpose devices (such as those used by delivery services such as Federal Express).
- Personal Digital Assistants (PDA) - such as PalmPilot or the Windows CE machines designed to run custom USDA programs or modified versions of off-the-shelf programs. CCE decided that PDAs for maintaining schedules, addresses and phone numbers were to be a personal preference item. Since the completion of the market survey, ESRI announced a new product, ArcPad, that when released in 2000, will provide GIS functionality on PalmPilot and Windows CE devices.

A market survey has been completed for the category of Sub-notebook Computers. An evaluation will be conducted in the near future for the remaining categories.

For the purposes of the market survey, sub-notebooks were defined as notebook computers under a 5 pound base weight with battery-life over 3 hours. Sub-notebooks have similar capabilities as a full-sized desktop or conventional laptop and run Windows 95, Windows 98, or Windows NT. Therefore, they can easily accomplish tasks such as the creation and manipulation of databases, custom forms, e-mail capabilities, and office automation.

After analyzing the USDA requirements and needs for sub-notebooks, the candidate list was narrowed to four sub-notebooks. The following chart (Exhibit 10) is a comparison of the requirements and the four units that satisfied the criteria:

Determining Requirements	IBM ThinkPad 570	Toshiba Protégé 3025	NEC Versa SX	Acer TravelMate 332T
Weight with Battery	4lbs, 6.9lbs with UltraBase (FDD/CD-ROM)	2.9lbs, 3.27lbs with port expander	4.8lbs, 5.4lbs with CD-ROM	4.1lbs, with battery
Battery Estimated Life / Type	3 hour Li-Ion	3.8 hour Li-Ion	3 hour Li-Ion with an option for a 3 hour with the VersaBay	3 hour Li-Ion
Internal Specifications	366MHz Mobile	300MHz Pentium MMX,	366MHz Mobile	366MHz Pentium II,

	Pentium II, 64/192 MB SDRAM, 6.4GB hard drive	96 MB EDO- DRAM, 6.4GB hard drive	Pentium II, 128/256MB SDRAM, 6.0GB hard drive +VersaBay	64/256 MB SDRAM, 4GB hard drive
Cost	Open Market Price: \$3499.00	GSA Price: \$1937.00	Open Market Price: \$3607.00	Open Market Price: \$2700.00

Exhibit 10: Top Sub-Notebook Candidates

Reference: Mobile Computing Market Survey Report, Draft, dated 8 June 1999.
Pp.1-2.

Appendix C: CCE Software Evaluations and Configurations

Workstation Database products tested as a result of the LTD effort were Microsoft Access, Corel (Borland) Paradox, and Lotus Approach. All three packages offer comparable basic database capabilities, such as creating a database with forms and reports. Access has the best user interface, followed by Paradox and then Approach. User friendliness seems to be a key strength of Access.

For cost efficiency purposes, it was decided to buy a single package suite to fulfill office automation functions. Based on the overall LTD results, which combined the scores for each functional product area, the Microsoft Office suite was selected for purchase for both the BPR pilots and the Y2K initiative.

Reference: USDA Service Center Business Need and Technical Alternative Evaluation Study – Phase II, dated 9 April 1998. Pp.75-106.

Groupware

Selection Status Summary: Microsoft Exchange was selected as the Groupware product for the BPR pilot sites. A final decision for national deployment has not been made.

Evaluation and Test Summary

Microsoft Exchange 5.5sp1 Enterprise Edition and Lotus Notes Domino 4.6.3 were tested as part of the USDA live test demonstration (LTD) effort.

Based on those requirements defined by USDA^{viii}, a preliminary market evaluation^{ix} was conducted that reduced the list of viable Groupware packages from four (Microsoft Exchange, Lotus Domino, Netscape SuiteSpot, and Novell GroupWise) to two (Microsoft Exchange and Lotus Domino). These remaining two software packages were tested between December 1998 and March 1999.

Generally, both product packages fulfilled the USDA requirements. Microsoft Exchange/Outlook excelled in ease of use, with both the client and the server following the Windows User Interface Guidelines closely, as well as providing rules wizards and predefined options for easy customization. Lotus Notes/Domino's user interface was not consistent with the Win32 User Interface Guidelines as well as requiring administration from both a Graphical User Interface (GUI) and a Command Line Interface (CLI). However, Lotus Notes/Domino excelled in complex customization, even providing sources for most of the templates and forms in the system, as well as supporting robust replication and remote access.

Final scores, both unweighted and weighted, were tied as the few discriminating categories between products were weighted equally. In the absence of clear prioritized discriminators in this category, Microsoft Exchange was selected for deployment in the BPR Pilot sites as it is bundled with Microsoft Back Office, which is currently deployed at all pilot sites. During the time Microsoft Exchange

is deployed at pilot sites, additional requirements may be identified. In addition, cost considerations at the time of procurement may be a deciding factor in making a final product selection.

References:

Groupware Testing Results, dated 1 June 1999. Pp. 1-2.

CCE Analysis and Testing Update, dated 3 June 1999. Pp. 2-3.

CCE MS Exchange Server Architecture and Implementation Plan for Service Center Pilot Sites, Draft, December 1999

Enterprise Geographic Information System Software

***Selection Status Summary:** ESRI ArcView Desktop GIS software was selected initially in early 1998 for use in the BPR pilot sites. In 1999, more extensive Enterprise level GIS Live Test Demonstrations also led to the selection of the ESRI suite of Enterprise GIS tools for initial deployment. On September 29, 1999, the Management Review Board (MRB) accepted the recommendation by the GIS team for an initial deployment of ESRI products to 450 Field Service Centers.*

Test and Evaluation Summary

During the fall of 1998, the USDA formed an Enterprise GIS Team to analyze GIS requirements and conduct market research to identify commercial off the shelf (COTS) products that best complied with USDA needs. COTS GIS products were specified to reduce the level of customization that may be needed to support USDA CCE GIS requirements.

A Test Plan Subteam of the Enterprise GIS Team defined a three-step process for GIS market research. Each step in this process was intended to identify the COTS GIS products that best met USDA needs and reduced the number of products under consideration for CCE. The following three steps were defined for the CCE GIS market research process: 1) Enterprise GIS Market Survey, 2) GIS Compliance Statement Review and 3) Live Test Demonstrations.

The first step, the Enterprise GIS Market Survey, was completed in February 1999. The following companies were evaluated in this study: Autodesk, Bentley Systems, ESRI, Genasys, Intergraph, MapInfo, MCI WorldCom VISION and SmallWorld. Companies which complied with over 85 percent of high level USDA requirements were selected for more detailed evaluation. Bentley Systems, ESRI, Intergraph, MCI WorldCom VISION and SmallWorld were selected to participate in the next phase of market research.

The second step, the GIS Compliance Statement Review, resulted in two companies, ESRI and SmallWorld, being selected to participate in intensive LTDs. ESRI received a score of 407.0 for the compliance statement review while SmallWorld's score was 399.5. The highest scoring GIS vendor in the LTDs was ESRI with a total score of 276.5. SmallWorld received a total score of 232.5.

ESRI products had the best out of the box functionality, better training support and user documentation and better DBMS support. SmallWorld offered better long transactions and history tracking capabilities.

Based on their better performance in the LTDs and superior out of the box functionality, ESRI products were recommended for initial Enterprise GIS deployment. Separate reports prepared by USDA GIS specialists and the consultants who scored the LTDs both concurred in the selection of the ESRI enterprise GIS product suite. On September 29, 1999, the Management Review Board (MRB) accepted the recommendation by the GIS team for an initial deployment of ESRI products to 450 Field Service Centers.

Reference:

USDA CCE GIS Results Report, dated 8 June 1999. Pp.1-3;

USDA Service Center Agencies Recommendation For National Enterprise; Geographic Information Systems (GIS) Solution, September 21, 1999

Database Management System (DBMS) Software

***DBMS Vision:** A single, homogeneous, common database environment at the Service Center level. Multiple, heterogeneous, databases addressing application-specific requirements at the region and agency levels.*

***Selection Status Summary:** No product has been selected. The evaluation based on Market Survey and Vendor Compliance Statements has been completed.*

Test and Evaluation Summary

IBM DB2, Informix Universal Server, Microsoft SQL Server, Oracle, and Sybase Adaptive Server were evaluated in the database management system (DBMS) category.

The DBMS market survey evaluation consisted of two parts:

- **Business Evaluation** - provided a look at the financial health of all five companies. The intent of the criteria was to confirm that the vendors have an adequate financial base, revenues, profits, growth rates and market share to continue developing and supporting enterprise DBMS products. Information was updated to reflect 1999 data.
- **Compliance Statement Evaluation** - where all vendors were asked to respond to a written list of USDA data management requirements. These compliance statements contained 157 requirements for various categories of DBMS systems to which the vendors were asked to respond and provide documentation for validation. This provided a functional and technical comparison of DBMS vendors.

The sources for the USDA DBMS requirements were two reports by the USDA. First was a working paper entitled “Service Center Implementation Team (SCIT) Data Management Tools Selection Strategy” dated February 1999, and second, the published revision of that report with a new title: “Service Center Implementation Team (SCIT) Data Management Tools: Requirements/Strategy” dated June 1999. The first working paper was used as the basis for the Vendor Compliance Matrixes that were completed by the vendors then evaluated and scored. The second paper, along with new financial data, is the primary reason for significant updates to this section.

In the absence of specific applications architecture and data distribution architecture, LTDs were not conducted at this time. LTDs are planned for the near future as part of a larger integration testing strategy.

The CCE Team added new requirements generated by the Data Management team, received additional input from vendors, and included the new requirements in the process to evaluate the vendors. The highest possible score that a vendor could receive was 471. Final recommendations are discussed below.

Oracle8 led the DBMS vendors in the compliance scoring by a mere 18 points, excelling in native support for GIS computing. **Exhibit 11** outlines the scores from the compliance statements and highlights some of the major points discovered in the business evaluation.

DBMS Product	Total Score (out of 471)	Summary
Oracle v8	451	Oracle is a likely candidate for selection by USDA due to its high compliance score and a strong showing in the marketplace. It led the field in 1997 Overall Market Share and for 1997-98 Market Share on UNIX and on NT. Can run GIS in native mode.
Sybase v11.5	433	Sybase had a very strong technical showing with 354 points, giving Sybase second place on compliance. But the company’s financial problems raise serious doubt as to their business/financial viability. Their shift in emphasis from product to services also raises doubt about their continued investment into their product line. Strong showing in mobile computing.
IBM DB2 v5.0 (Unix/NT)	425	If USDA CCE uses AS/400 for any of their database platforms, then IBM’s DB2 is the only choice for a DBMS on that platform. IBM financially is solid. On the UNIX/NT platforms, IBM placed third in the compliance scores. On the AS/400 platform it placed fifth

The reason for this is that specific applications and platforms are only now being identified for the Service Center Initiative. Each of these projects will have specific business and functional requirements, which will in turn, translate to more specific technical database requirements.

Any final decisions should include other factors, such as: application-specific requirements; development and delivery platforms; number of users; and live testing with USDA data on the proposed platforms. In addition, decisions should also be based on the financial health of the company and the vendor's ability to support USDA in the short and long term.

Reference: Database Management System (DBMS) Market Survey and Functional Capabilities Report, Draft, dated 30 November 1999.

Document Management System (DMS) Software

Selection Status Summary: No product has been selected. The evaluation based on Market Survey and Vendor Compliance Statements has been completed.

Test and Evaluation Summary

This market survey compared the business and financial viability of alternative vendors of Document Management Systems (DMS) and compared the high-level functional and technical capabilities of their software products to Common Computing Environment (CCE) requirements for DMS technology. This report described the comparison of DMS technology and USDA requirements to be implemented at the Service Center level, State/Regional level or at the Centralized Architecture level (1-4 physical locations to support nationwide requirements). Throughout USDA documents (e.g., Benefit Cost Analyses, Technical Requirements, Technical Architectures, and other documents), a variety of terms are used for what appears to be the same thing. These terms include: "Imaging," "Workflow," "Version Management," and others. For the purpose of this survey, the term "Document Management Systems" or "DMS" was used, and it included several different aspects of enterprise document management such as document imaging, document versioning, workflow, document viewing, document storage, reporting, and record management.

Summary of Requirements

The chart below summarizes the USDA DMS business and technical requirements collected to date. DMS functions are listed in order to compare the requirements to DMS capabilities.

DOCUMENT MANAGEMENT SYSTEM SOFTWARE FUNCTION	USDA REQUIREMENTS
Document Imaging	10,000+ pages a day at Development Centers; storage of archived documents for the Service Center Initiative (SCI) Project

DOCUMENT MANAGEMENT SYSTEM FUNCTION	SOFTWARE	USDA REQUIREMENTS
Document Management: Check-In/Check/Out; Versioning; Access control; Indexing; Profiling		Management of all customer related documents created, processed, and routed for approval
Document Security		Security needed for confidential loan processing and all documents used
Integration with Other Applications		MS Office; MS Exchange; MS Outlook; ESRI GIS Products; RDBMS; Enterprise Web Server; Enterprise Management System software; etc.
Interface Compliance		Interface with local community or government systems and applicable program information
Workflow Management: Ad-hoc Workflow; Message Routing; Administrative Forms		USDA foresees starting with simple routing of loan applications for approval and moving eventually to complex tracking of loan processing and other internal functions, such as debt tracking; access to and routing of a variety of forms
Remote Access		Secure access to multiple document types from remote computer equipped vans or other rural locations
WEB Enabled		WEB access to USDA program information; loan status; +other applicable documents by internal staff and customers; future loan application on-line
Configuration Management		Track changes to land owned by USDA customers and associated changes to related loans
Operating System		Windows NT 4.0
System Security		NT rule-based security
API or Server Strategy		Customize system to integrate with Microsoft products in use
Scalability		Meet Service Center to National level requirements
Flexibility		Multiple platforms, variety of databases, distributed technology
Company Focus		General office documents; and access to multiple document types including maps
Complexity		Ease of install and use for low deployment and support costs

Exhibit 12: USDA DMS Requirements Summary

DMS Summary

This survey researched and then narrowed the field of DMS vendors to three market leaders with the greatest overall market share today. They include FileNET, Documentum, and Open Text. The strengths of the three vendors include:

- **FileNET** is widely recognized as the market leader in **document imaging services**. The latest release of its Panagon suite of products is expected to continue where prior versions left off. FileNET, through several acquisitions of leading DMS niche

players, has developed an integrated suite of DMS products to provide an end-to-end document management solution.

- **Documentum** is currently recognized as one of the market leaders in the **document management market**. By providing a suite of integrated DM tools, Documentum provides a solid end-to-end solution. Although imaging is not a Documentum product line, the company has 3rd party partnerships to provide imaging solutions.
- During the last two years, no other DMS vendor has enjoyed as much growth as **Open Text**. Open Text is recognized as the leading provider of **web-based document management**. The product is built from the ground up on web technology and standards. Although imaging is not an integrated product line for Open Text as well, the company also has 3rd party partnerships to provide imaging solutions.

Recommendations

- USDA should select a market leader who adheres to the developing DMS standards for its enterprise-wide DMS solution.
- All three of the selected vendors discussed in this paper are capable of providing an enterprise-wide DMS solution.
- It is suggested that the vendors provide product demonstrations to appropriate USDA staff, that USDA further define its requirements, and then the top requirements be prioritized in order to determine which vendor could provide the greatest benefit for the highest priority needs. Also, all three vendors should be encouraged to present their enterprise-wide solution and product lines that meet all requirements.
- CCE requirements for imaging documents at Field Service Centers are non-existent currently, but imaging is used at Partner Agency Development Centers. Over time and as more document input becomes Web-based, the requirements for imaging should diminish. This will take time, however, so some imaging capabilities will need to be provided in the interim.
- USDA is already piloting FileNET with one BPR project. After priorities are documented, the pilot testing of the other two vendors for other BPR areas may be warranted, as long as funding is available.
- USDA should stay abreast of developments with the bundling of DMS capabilities into Lotus and Microsoft product suites. Some of the projected functionality for these products may meet some USDA requirements and be provided at little or no cost. In terms of Lotus products, the Lotus/IBM DMS solutions would be cost effective if USDA decided to purchase Lotus as its groupware provider. Otherwise, in order to use Domino.Doc and other Lotus-based DMS products, USDA would have to purchase Lotus products to make the DMS technology work. IBM and its Lotus Domino products were eliminated as not cost-effective now; this could change in the future if USDA selects Lotus as its groupware solution.
- A DMS implementation for USDA is not a simple undertaking. Customization would be required and users would need to be trained in new ways of conducting their work.

As USDA prepares for further pilot projects using DMS products, the items discussed above should be considered. If funding constraints prohibit further

review and analysis of DMS products, another survey of the market closer to an implementation date may be needed in order to update potential users on new technological capabilities and the entry or exit of key market players.

Reference: CCE Document Management Systems, Market Survey Report, Draft dated November 15, 1999.

Appendix D: CCE Transaction Processing Requirements

A.5.1 Definition

Transaction processing (TP) services provide support for the on-line processing of information in discrete units called transactions, with assurance of the state of information at the end of the transaction. This typically involves predetermined sequences of data entry, validation, display, and update or inquiry against a file or database. It also includes services to prioritize and track transactions. TP services may include support for distribution of transactions to a combination of local and remote processors.

A.5.2 Requirements

FSC Requirement
Reliability of data in a distributed environment

A.5.3 Enabled Services

Typically, a transaction processing service contains a transaction manager, which links data entry and display software with processing, database, and other resources to form the complete service. The sum of the work done anywhere in the system in the course of a single transaction is called a global transaction. Services associated with the Transaction and Resource Managers are described below:

- Transaction Manager
 - Starts a transaction
 - Opens and closes resource managers
 - Commits or rolls back transactions
 - Chains transactions together
 - Monitors transaction status
- Resource Manager
 - Provides access to shared resources such as databases, file access systems or communications facilities

A.5.4 Importance of Element

The significance of the transaction processing services is they allow the following:

- Increased efficiency of shared resources,
- Access to information independent of location,
- Improved data consistency and accuracy.

A.5.5 Applicable Standards

- Protocol for heterogeneous interoperability - ISO 10026-1,2,3:1992 (OSI Distributed Transaction Processing)
- Transaction manager-resource manager interface - X/Open C193:1992 (XA Specification)
- Transaction demarcation - X/Open P209:1992 (TX Specification)
- Transaction manager to communications manager interface;
X/Open S423:1994 (XA+ Specification)
X/Open P306:1993 (XATMI Specification)
X/Open P306:1993 (TxRPC Specification)
- Distributed queuing - IEEE P1003.15 (POSIX Batch Extensions)

A.6 Appendix F: CCE Distributed Computing (Tier I)

A.6.1 Definition

Distributed computing services provide specialized support for applications that are physically dispersed across a number of application platforms yet are maintained in a cooperative processing environment. The classical definition of a computer becomes blurred as the processes that contribute to information processing become distributed across a facility or a network.

A.6.2 Requirements

FSC Requirement
Minimal technical support required <i>in</i> the FSC
Automated roll-up of management level reporting data
Quick access to data

A.6.3 Enabled Services

The distributed computing element enables the following services:

- Distributed Time,
- Distributed Data,
- Distributed File,
- Distributed Name,
- Remote Processing, and
- Remote Print Spooling and Output Distribution.

A.6.4 Importance of Element

The significance of distributed computing standards is that they allow the following:

- Access to information independent of location,
- Scalability and fault tolerance,
- Increased efficiency of shared resources,
- Support of reorganization, and
- Remote systems support.

A.6.5 Applicable Standards

- ISO/IEC 9636-1..6:1991 (CGI) Device interfaces - Device interface API
- OSF DCE 1.1: DFS Distributed computing environment services - Distributed file service
- OSF DCE 1.1: DTS Distributed computing environment services - Distributed timing service
- OSF DCE 1.1 Cell Directory Service/ Global Directory Service Distributed computing environment services - Naming services
- OSF DCE 1.1: RPC Distributed computing environment services - Remote procedure call

Appendix G: CCE Geographic Information Systems (GIS) (Tier II) Requirements

A.10.1 Definition

Within this section, GIS is defined as the software, data, and procedures used to acquire, store, manage, analyze, view, and print geographic data. Geospatial and associated attribute data are referenced to the surface of the earth through a coordinate system.

A.10.2 Requirements

FSC Requirement
Utilize geospatial software to hasten analysis and automate currently manual processes
Utilize a USDA-wide base set of geospatial data
Each FSC is supplied with GPS units

A.10.3 Enabled Services

The Geographic Information Systems element consists of the following services:

- Graphical Object Management (to include geospatially referenced data),
- Drawing, and
- Imaging.

A.10.4 Importance of Element

Many USDA business activities occur in a specific location, particularly for the land management and FSC agencies. The relationship of natural and cultural resources and the impacts of alternative management practices on them can best be analyzed and portrayed geospatially. Maps are very powerful means of conveying a maximum amount of information on topics as diverse as animal damage assessments, forest management plans, risk management studies, and conservation plans. Tight integration between geoprocessing functions and database management systems is required for effective use of GIS within USDA.

The Open Geodata Interoperability Specification (OGIS), is "a comprehensive specification of a software framework for distributed access to geodata and geoprocessing resources. OGIS will give software developers around the world a

detailed common interface template for writing software that will interoperate with other OGIS compliant software written by other software developers".^x

USDA/FSC supports the efforts and direction of the OGIS standard, has representation at the technical committee and management committee levels, and will adopt interim versions as well as the final standard.

A.10.5 Applicable Standards

- OGIS
- Spatial Data Transfer Standard (SDTS), FIPS PUB 173-1
- Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata
- Vector Graphics Data, FIPS Pub 128 (CGM)
- Raster Data Interchange:
 - NIST FIPS PUB 150 (Group 4 Facsimile)
 - NIST FIPS PUB 158-1 (X-Windows, for BDF)
- Still Image Compression:
 - NIST FIPS PUB 147 (Group 3 Compression)
 - NIST FIPS PUB 148 (General Facsimile)
 - NIST FIPS PUB 150 (Group 4 Facsimile)
 - ITU-T T.81-1993 (JPEG)
 - ISO/IEC 10918-1 (JPEG)

Appendix H: CCE Security Requirements (Tier III)

A.11.1 Definition

The USDA security policy defines the relevant security requirements and measures (including standards) that different platforms must implement to create a secured network infrastructure. FSC will strictly comply with this policy. The policy addresses the full spectrum of security needs, including confidentiality, integrity, and availability. Confidentiality requirements protect against inappropriate disclosure of information; integrity requirements ensure the correctness and appropriateness of information and/or its sources; and availability ensures that information is present and usable within reasonable time constraints.

In addition, the security policy includes the internal security controls (technical security measures) that are implemented in hardware, firmware, and software of automated information systems (AIS). In order for internal security controls to be effective, adequate external security controls, which include physical, personnel, procedural and administrative security measures, will be employed. These security measures are the foundation upon which all other security should be built.

Today, almost every computer is electronically connected to other computers across multiple platforms through modems, LANs, WANs, or the Internet. For this reason, the above mentioned security measures need to be implemented. The security of a network is only as good as the weakest link in the security chain, whether it is administrative, personnel, or a technical control.

It is important to develop security implementation in a layered approach, with each layer complementing the next. With each additional layer, the ability to access, modify, or destroy AIS resources or facilities becomes geometrically more difficult. For example, effective password controls (administrative) may be level 1. Add the use of smart cards (technical) as layer 2. Add the use of alarm systems (physical) as layer 3, etc. With each added layer, the likelihood of successful, unauthorized access is significantly reduced and the network, as a whole, becomes increasingly more secure.

A.11.2 Requirements

FSC Requirement
Confidential customer data must be secured
FSC staff must have access to all Partner Agency systems
Database systems must be capable of

supporting/enforcing varying security levels
Public vs. Confidential data must be delineated prior to Web publishing
Systems must be protected from viruses

A.11.3 Enabled Services

Counters to security threats are provided by the following services:

- Identification and Authentication Services,
- System Entry Control Services,
- Audit,
- Access Control,
- Non-Repudiation Services,
- Security Management Services,
- Trusted Recovery Services,
- Trusted Communication Services (including encryption), and
- Anti-Virus Protection.

A.11.4 Importance of Element

The purpose of security is to:

- Assure compliance with security laws and regulations,
- Protect from virus, physical and electronic intrusion, or disaster,
- Minimize risk to information assets, and
- Assure appropriate access to external information supplied to the government.

A.11.5 Applicable Standards

- NIST FIPS PUB 112 (Password Usage) Architectures and applications - Operating system security
- NIST FIPS PUB 113 (Computer Data Authentication) Authentication
- NIST FIPS PUB 140-1 (Security Requirements for Cryptographic Modules) Confidentiality - Data encryption security
- NIST FIPS PUB 185 (EES) Confidentiality - Data encryption security
- NIST FIPS PUB 46-2 (DES) Confidentiality - Data encryption security
- NIST FIPS PUB 74 (Guidelines for DES) Confidentiality - Data encryption security
- NIST FIPS PUB 81 (DES Modes of Operation) Confidentiality - Data encryption security

- PL 100-235 (Computer Security Act of 1987) Confidentiality - Open systems confidentiality
- PL 93-579 (Privacy Act of 1974) Confidentiality - Open systems confidentiality
- FIPS PUB (DSS)* DRAFT Digital Signature
- IEEE 1003.1b:1993 (POSIX Real-Time Extensions) System management security - Security Management
- NIST FIPS PUB 151-2 (POSIX.1) System management security - Security management
- NIST FIPS PUB 191 (Guideline for LAN Security)
- Computer Security Act of 1987 (Public Law 100-235)
- Computer Fraud and Abuse Act of 1986 (Public Law 99-474)
- Freedom of Information Act of 1980 (Public Law 93-502)
- Federal Managers' Financial Integrity Act of 1982 (Public Law 97-225)
- Electronic Communications Privacy Act of 1986 (Public Law 99-508)
- Privacy Act of 1974 (Public Law 93-579, 5 United States Code 552a, July 14, 1987)
- Executive Order 10450 of April 27, 1954
- Federal Personnel Manual (FPM), Chapter 736-13, 1988
- Copyright Act (17 United States Code 105)
- U.S. Office of Government Ethics, Standards of Ethical Conduct for Employees of the Executive Branch
- OMB Circular A-130, Management of Federal Information Resources, Appendix III, Security of Federal Automated Information Resources, February 8, 1996
- OMB Circular A-123, Management Accountability and Control, June 21, 1995
- OMB Circular A-127, Financial Management Systems, July 30, 1993
- OMB Bulletin 90-08, Guidance for Preparation of Security Plans for Federal Computer Systems That Contain Sensitive Information, July 9, 1990
- USDA Departmental Regulation (DR) 3140-1, USDA Information Systems Security Policy
- USDA Departmental Manual (DM) 3140-1, USDA Information Systems Security Manual
- USDA DR 3140-2, USDA Internet Security Policy
- USDA DR 3300-1, USDA Telecommunications, Section 4, Appendix I
- USDA Employee Handbook, Appendix I, Employee Responsibility and Conduct
- USDA DR 3130-2, USDA Microcomputer Policy
- USDA DM 3440-1, USDA Classification, Declassification, and Safeguarding Classified Information
- Federal Information Processing Standards (FIPS) and National Institute of Standards and Technology (NIST) Special Publications

Appendix I: CCE Interfaces Requirements (Tier III)

A.12.1 Definition

Application interfaces provide a “common ground” facility so that interaction between different applications may occur with minimal impact on the user.

The list below provides examples of areas where application interfaces allow agencies to share information:

- Office Automation (OA) software including spreadsheets, graphics, word processing, and mail,
- Geo-spatial data,
- Multimedia,
- Compatibility with graphic format exchange, and
- Middleware to enable non-direct connectivity of systems.

Unfortunately, incompatible versions of the same system software are currently prohibiting the ability to share information among agencies.

A.12.2 Requirements

FSC Requirement
Application interfaces to shared customer database
Application/Database interfaces to geospatially stored data
COTS application interfaces available

A.12.3 Enabled Service

Interfaces serve to provide interoperability between application environments.

A.12.4 Importance of Element

The significance of application interfaces is that they:

- Improve exchange of information internally and externally (Other FSC, Agencies, Departments, State/Local Governments, citizens/customers, trade associations, advocacy groups, and others),
- Update data quickly,
- Improve collaboration between mission and external groups,
- Increase accuracy through data reuse,
- Involve time efficiency with established pathway of exchange, and

- Improve data consistency.

A.12.5 Applicable Standards

- NIST FIPS PUB 173-1 (Spatial Data Transfer Standard) Geospatial data exchange
- NIST FIPS PUB 1-2 (Code for Information Interchange) Characters and symbols - Character sets
- ISO 11172-1,2,3:1993 (MPEG) Compression - Motion image compression
- ISO/IEC 10918-1 (JPEG) Compression - Motion image compression
- X/Open C436:1994 (Commands and Utilities) Compression - Text and data compression
- NIST FIPS PUB 152 (SGML) Document interchange - Custom definition of document types
- NIST FIPS PUB 152 (SGML) Document interchange - Document exchange
- NIST FIPS PUB 177 (IGES) Technical data interchange - Vector graphics data interchange
- ISO/IEC 9592-4:1992 (PHIGS PLUS) Vector graphics - Vector graphics API
- NIST FIPS PUB 153 (PHIGS) Vector graphics -Vector graphics API
- NIST FIPS PUB 128-1 (CGM) Vector graphics -Vector graphics data interchange

Appendix J: CCE System and Network Management Requirements (Tier III)

A.13.1 Definition

Information systems are composed of a wide variety of diverse resources that must be managed effectively to achieve the goals of an open system environment. The basic concepts of management, including operation, administration, and maintenance may be applied to the full suite of information system components along with their attendant services.

A.13.2 Requirements

FSC Requirement
Minimal IT support strain on FSC personnel

A.13.3 Enabled Services

System and Network management services include:

- Configuration Management Services,
- Performance Management,
- Availability and Fault Management,
- Accounting Management,
- Security Management,
- Print Management Services,
- Network Management,
- Backup and Restore,
- On-line Disk Management,
- Software and Hardware Inventory Management,
- Capacity Management, and
- Software Installation.

A.13.4 Importance of Element

The significance of system and network management is that they:

- Enable remote administration,
- Provide information on what the user is doing and how much it costs,
- Increase reliability,
- Assure consistency of procedures,

- Maximize efficiency of resources,
- Enhance problem diagnostics, both locally and remotely,
- Reduce down time,
- Increase flexibility in configuration management,
- Provide data for analysis and planning on operations,
- Establish performance measurement,
- Enable capacity planning and performance management, and
- Protect information resources from unauthorized access.

A.13.5 Applicable Standards

- SNMP II

End Notes

ⁱ Adapted from NRCS web site at <http://www.ncg.nrcs.usda.gov/who.html>, last updated March 24, 1997.

ⁱⁱ Adapted from RD web site at <http://www.rurdev.usda.gov/programs.html>, last updated March 17, 1997.

ⁱⁱⁱ Farm Service Agency Draft Strategic Plan – Fiscal Years 1997 - 2002, p. 2. As of December 1996, FSA was located in 2386 Service Centers, NRCS in 2493, and RD in 1287.

^φ Reprint from the USDA, Common Computing Environment, Information Technology Architecture Version 1.1 Draft. December 10, 1999

^{iv} USDA SCI Modernization Plan, November 1999

^v From the CITRIX web site, Server-based Computing White Paper, December 1999

^{vi} *Data Administration Concept of Operations, Glossary*, USDA, June 15, 1998

^{vii} *Data Administration Concept of Operations*, USDA, June 15, 1998. Pg 22

^{viii} From the USDA IOP Memorandum detailing the USDA Groupware Requirements and from additional Partner Agency Contact specified requirements.

^{ix} The market evaluation is documented in the September CCE Analysis & Testing Update on the Joint Discussion database.

^x *The OpenGIS™ Guide*, ed. Kurt Buehler and Lance McKee (Wayland, Massachusetts: Open GIS Consortium, Inc, 1996), P4.