



## Architectural Design and Best Practices Project Final Report and Design Recommendations (A006.1)

Prepared for the Virginia Department of Education  
February 28, 2011



**Technical Point of Contact:**

Louis McDonald | CIO/CTO

[louis.mcdonald@cit.org](mailto:louis.mcdonald@cit.org) | 703.689.3037

---

**Administrative Point of Contact:**

Pat Inman | Contract Manager

[pat.inman@cit.org](mailto:pat.inman@cit.org) | 703.689.3037

## Contents

Executive Summary .....	4
1 Introduction .....	6
1.1 Study Goal .....	6
1.2 Project Deliverables .....	6
2 Research Process .....	6
3 Key Messages .....	8
3.1 Stakeholder Management .....	8
3.2 Federated Systems Perform Poorly .....	9
3.3 Data Governance .....	9
3.4 Leveraging Existing Systems .....	10
3.5 Use of Commercial Systems .....	10
3.6 Multiple Hash Keys .....	11
3.7 Requirements and System Architecture .....	11
3.8 Clearly Defined Security Policies .....	12
4 Architecture Best Practice Case Studies .....	13
4.1 Indiana Department of Education .....	13
4.2 Iowa Department of Education .....	18
4.3 Army Suicide Mitigation Project .....	20
4.4 Texas Education Agency .....	23
4.5 DLA Data Convergence and Quality Project .....	26
4.6 NORC at the University of Chicago Data Enclave .....	29
5 Subject Matter Expert Interviews .....	34
5.1 Dr. Bhavani Thuraisingham .....	35
5.2 Paul Carney .....	36
5.3 James Campbell .....	37
5.4 Susan Carter .....	38
5.5 Raj Ramesh .....	40
5.6 Ron Kleinman .....	41
5.7 Peter Dobler .....	42
5.8 Dr. Laura Haas .....	43
5.9 Dr. Thilini Ariyachandra .....	44
5.10 Dr. Cynthia Dwork .....	45
6 SLDS Architecture Overview .....	47

6.1	SLDS Seven Functional Components .....	47
6.1.1	Portal.....	47
6.1.2	Security .....	49
6.1.3	Workflow .....	50
6.1.4	Reporting.....	53
6.1.5	Lexicon.....	56
6.1.6	Shaker.....	57
6.1.7	Data.....	58
7	Physical Infrastructure .....	60
7.1	Development Environment.....	60
7.2	Test Environment.....	61
7.3	Production Environment.....	62
Appendix A: Secondary Architecture Best Practice Case Studies.....		63
A.1	Illinois State Board of Education .....	63
A.2	North Dakota Department of Public Instruction.....	65
A.3	Washington Education Research and Data Center.....	68
Appendix B: Best Practice Case Studies Interviewee List .....		70
B.1	Indiana Department of Education.....	70
B.2	Iowa Department of Education.....	70
B.3	Data Strategies – Army Suicide Mitigation Project .....	70
B.4	Texas Education Agency .....	71
B.5	Data Strategies – DLA Data Convergence and Quality Project .....	71
B.6	NORC Data Enclave .....	71
B.7	Illinois State Board of Education .....	72
B.8	North Dakota Department of Public Instruction.....	72
B.9	State of Washington Education Research & Data Center .....	73
Appendix C: Materials Sent to Best Practices Interviewees.....		74
C.1	Best Practices Interview Template.....	74
C.2	Architectural Best Practice, Design & Planning Support Project Overview .....	76
Appendix D: Materials Sent to Subject Matter Experts .....		77
D.1	Subject Matter Expert – Interview Template .....	77
D.2	Virginia Statewide Longitudinal Data System - Executive Summary .....	79
D.3	Virginia Statewide Longitudinal Data System - Usage .....	82

## Executive Summary

The educational landscape has changed dramatically since the establishment of Statewide Longitudinal Data Systems (SLDS) throughout the United States. A grant program<sup>1</sup> funded by the U.S. Department of Education, as authorized by the Educational Technical Assistance Act of 2002, has helped to change many states' K-12 data systems significantly and may, in fact, revolutionize the future management and utility of educational data. States that have implemented these data systems now have more accurate and robust data and an enhanced ability to access, analyze and utilize the data in a manner previously unavailable in the past. These successes, however, have been achieved only through rigorous planning, meticulous design, and assiduous implementations.

In 2010, the Virginia Department of Education (VDOE) was awarded a federal multi-year grant to enhance its statewide data system, and launched the Virginia Longitudinal Data System (VLDS) project. The project team was charged with creating a system that will address the needs of all stakeholders; additionally, the team faced challenging Federal and Virginia security and privacy requirements. To meet these challenges, VDOE commissioned the Center for Innovative Technology (CIT) to identify key success factors that could provide guidance in the development of a fully secure and private, as well as more efficient, SLDS in Virginia.

In order to achieve this complex objective, the CIT project team investigated other large data integration projects in state education agencies, other governmental agencies and in other industries. The team examined published information from a variety of sources, and a gap analysis was performed on these reports and articles to identify missing information and critical areas where further research must be conducted. In addition, the team interviewed nine SLDS leaders and a number of industry leaders who managed large integration projects. These interviews provided the foundation for the compilation of best practices and key takeaways that were essential to the success of each of these projects.

Information collected during the research and analysis phase was analyzed to identify common themes and these themes were organized into a set of best practices that were included in a final report. This process revealed several key elements<sup>2</sup> that were crucial to ensuring successful implementations. The primary themes that emerged included, among others, the necessity for detailed project planning and management (the importance of data governance and stakeholder management) as well as the need to conduct comprehensive research and planning before implementing the technology and creating the system architecture (the use of commercial solutions and leveraging existing systems).

---

<sup>1</sup> The Statewide Longitudinal Data Systems (SLDS) Grant Program

<sup>2</sup> The final report included challenges and obstacles to be avoided and provided recommendations for a number of preliminary action items. In addition, the report contained a supplement that included the complete reports with detailed findings for each interview and data project researched.

The CIT project team, along with VLDS subject matter experts, took the original conceptual SLDS architecture and incorporated results from the best practices research and subject matter expert interviews to develop an implementation architecture. The refined architecture (Section 6) expanded the details for functional components, security, reporting, and how data requests would be managed given known constraints. The team developed an understanding of the flow of information, and the necessary workflows to support the different scenarios to support the SLDS deployment.

All this information allowed the team to develop a physical infrastructure architecture (Section 7). This included the physical hardware, the location, and the functionality for the hardware. Following standard lifecycle development practices, three versions of the infrastructure are described; Development, Test, and Production.

## 1 Introduction

### 1.1 Study Goal

The goal of the Architectural Design and Best Practices Project was to provide the Virginia Department of Education (VDOE) with an up-to-date and relevant assessment of the best practices related to the design, development, deployment, and operation of a Statewide Longitudinal Data System (SLDS).

### 1.2 Project Deliverables

Deliverable ID	Description	Delivery Date
A001	Monthly Status Reports	Monthly
A002.1	List of all SMES interested in participating in interview program	October 2010
A002.2	Interview Schedule and Interview Template	October 2010
A002.3	Preparation and pre-reading material as required for Interview SMEs	October 2010
A002.4	Consolidated output from interviews	December 2010
A002.5	Template for Final Deliverable	November 2010
A003.1	Monthly Status Reports	Monthly
A003.2	PMO Support, to include program documentation i.e., Work plan, Scope, Requirements, Schedule, Risk, and Change Management plans as requested by the Program Office	Ongoing
A004	Architectural Best Practices Report	December 2010
A005.1	Workshop Agenda and Presentation material	December 2010
A005.2	Summary Workshop Findings Report	January 2011
A006.1	Final Report & Design Recommendations	February 2011

## 2 Research Process

The CIT project team applied its established CIT Connect research and analysis process to execute this effort. This rigorous “best practice” methodology includes the identification of and analysis of information provided by Subject Matter Experts and by comparable SLDS projects or large data integrations from the public and private sector provides high confidence results. Additionally, this method also includes analysis and consolidation of feedback received from the SLDS stakeholders.

**Figure 1: CIT Connect Process**



CIT Connect projects are performed under the control of a well-defined project management approach. This approach provides visibility into project status at all times via regular reviews, status reports, and interim deliverables. The CIT Connect Process, shown in Figure 1, provides a diverse, five-step approach to sourcing innovation, rigorous analysis of alternatives, and a

structured engineering methodology for creating final recommendations. The use of this structured multi-step approach maximizes the likelihood of success, while reducing risk. Project execution is guided by a project plan developed and maintained by the CIT project management team.

For the VDOE Architectural Design and Best Practices Project, the process began with an analysis of the information and requirements provided by VDOE and the identification of the information and technology areas to be targeted for study. The second process step focused on initial resource sourcing by developing a list of candidate SLDS and large data integration projects and Subject Matter Experts in the information and technology areas defined for the project. This second process produced several deliverables focused on Subject Matter Experts interviews and best practices case studies that were presented to the Virginia Department of Education between October 2010 and December 2010.

The Architectural Best Practices Report (A004) focused on initial data sourcing, both to develop a list of candidate SLDS and large data integration projects to be analyzed and to collect published analyses, reports, and case studies on existing projects. The set of candidate projects, which was documented in Deliverable A002-2, was compiled by accessing CIT's business network and by searching our data resources to identify candidate companies and organizations. This list was delivered to the Virginia Department of Education on October 31, 2010.

The Consolidated Output from Subject Matter Expert Interviews (A002.4) focused on initial resource sourcing by developing a list of candidate Subject Matter Experts in the information and technology areas defined for the project. CIT researchers reached out to CIT's business network in addition to contacting these identified experts, who were discovered through literature searches. The set of candidate Subject Matter Experts, which was documented in Deliverable A002.1, was submitted to the Department of Education on October 31, 2010.

For both reports, CIT reviewed the sourced materials and developed a gap analysis of information requirements. The gap analysis was the basis for the creation of survey tools and an analysis framework which then were used to guide the interview process and the assessment steps that followed. The gap analysis also guided the selection of leaders of projects targeted for direct interviews.<sup>3</sup> The information priorities that emerged from the gap analysis were, therefore, used to classify and to prioritize interview candidates via a process based upon criteria that included domain relevance to VDOE, cost, complexity, technical and business maturity, and stakeholder considerations. The survey questions and targeted list of candidates for best practice interviews were provided to the Department of Education as Deliverable A002-2 on October 31, 2010. The targeted list of Subject Matter Experts (Deliverable A002.1), interview schedule, question template (Deliverable A002.2), and pre-reading material (Deliverable A002.3) were provided to the Department of Education on October 31, 2010.

The subsequent phases of the process focused on consulting leaders of similar state and commercial efforts, synthesizing and analyzing and organizing their feedback and presenting these best practices to stakeholders. Step three, the longest phase of the project, involved

---

<sup>3</sup> All of the projects had potential relevance to the effort; however, it would have been both infeasible and duplicative to interview the leaders of all of the candidate projects, given this project's short timeline.

performing interviews and synthesizing and analyzing the information from both the Subject Matter Experts and Best Practices candidates. During this phase, CIT researchers conducted nine interviews with leaders of large-scale projects representing both the public and private sectors,<sup>4</sup> and ten Subject Matter Experts.

In step four, the CIT Connect team organized and categorized key lessons learned from each of the interviews and highlighted common themes and unique insights. The CIT Connect team presented the best practice and Subject Matter Expert interview findings to VDOE stakeholders on December 13, 2010 and integrated feedback and guidance from the stakeholders to generate the reports A004 and A002-4.

Step five involved the presentation of the CIT project team’s implementation architecture and associated physical infrastructure to the SLDS stakeholders on January 27, 2011. Presentations were made by the CIT project team and VLDS subject matter experts. The purpose of this workshop was to formalize an agreed upon architecture for the SLDS.

### 3 Key Messages

The three components of the VDOE Architectural Design and Best Practices project covered three different themes and focus. The Best Practices interviews and analysis centered on the implementation and logistical processes involved in large scale data integration projects. The Subject Matter Experts interviews focused on technical best practices for an LDS architectural design.

The following topics are the key messages and best practices borne out of the Best Practices and Subject Matter Interviews.

Best Practices Interviews	Subject Matter Experts Interview
Stakeholder Management	Federated Systems Perform Poorly
Data Governance	Data Governance
Use of Commercial Solutions	Use of Commercial Solutions
Leveraging Existing Systems	Use of Multiple Hash Keys
Requirements Drive System Architecture	Clearly Defined Security Policies

#### 3.1 Stakeholder Management

When embarking upon a systems integration project, numerous stakeholders play a part in the planning, development, implementation, and maintenance of the system. Knowing stakeholders’ requirements, expectations, and resources are essential to a project’s success.

During the Army Suicide Mitigation Project, Data Strategies discovered that managing the stakeholders became an overwhelming task when it came to obtaining memorandums of understanding (MOU) and data sharing agreements needed prior to the integration of a data source into the system. Data Strategies also found that clear communication between the project implementation team and the stakeholders as well as communication among the stakeholders was best facilitated by the project managers. The DLA Data Convergence and Quality Project

<sup>4</sup> A list of the organizations and companies with whom we spoke may be found in Appendix B.

managers needed to ensure accurate and timely communication of project status, feedback, and next steps; this created a foundation for a positive collaborative environment. This positive collaborative environment amongst the teams contributed to the overall success of the project.

The Indiana Department of Education attempted to gather stakeholder requirements through large, monthly meetings before discovering that meetings with individual stakeholder groups proved to be the more effective and led to increased buy-in.

Lastly, as Illinois is currently in the design stage of its SLDS, the Illinois State Board of Education has hired a consulting firm to perform some stakeholder management. The consulting group is in the midst of gathering the technical and program information for each of the 13 data systems that will be integrated into the Illinois SLDS and this information will have a direct impact on the SLDS' final architecture.

*(Army Suicide Mitigation, DLA, Illinois State Board of Education, Indiana Department of Education)*

### **3.2 Federated Systems Perform Poorly**

Federated systems suffer in performance more than a centralized database. Requirements and queries should be planned prior to building the system to maximize performance. Usually with a distributed database model, data converges into a warehouse for simpler analysis. A global schema is defined that allows for an easier convergence of the data. The federated model for the SLDS cannot allow for permanent convergence, nor is it likely that a global schema will be developed that encompasses all data sources. The disparate data and network impacts on the distribution of the data can impact the overall performance of the federated architecture.

*(Ariyachandra, Dobler, Haas)*

### **3.3 Data Governance**

The importance of data governance was a common message throughout the course of this effort. Data governance often is viewed as a large initial effort for many data integration projects. For systems that continue to expand and to add data sources, however, it will be an ongoing effort, one that, our interviewees noted, is often underestimated. In most cases, an SLDS effort must accommodate a number of disparate stakeholders and sources and, thus, requires a higher than normal level of effort to identify data ownership and oversight to ensure its accuracy and security. The fact that each data source will have its own data governance creates an additional layer of complexity when attempting to create and manage data governance. Prior to implementing a statewide system, it is critical for the stakeholders to agree on who owns the data in the system, who will oversee and maintain the system and who will approve output and requests for access.

The North Dakota Department of Public Instruction and the Washington Research and Data Center are two projects that are in the early stages of building their data integration systems. Both of these projects have had difficulties moving forward with the implementation of the system due because of stakeholders' inability to agree upon the rules of the SLDS' data governance.

The Army Suicide Mitigation and the DLA Data Convergence and Quality Projects were projects that stressed the importance of data governance during the early planning stages prior to implementation and continue to emphasize these elements as data sources are added. With the

addition of a new data source, the project managers must understand the new source's governance and how its integration will affect the LDS' overarching governance.

Appropriate data governance, particularly in a federated model, is key to ensuring that the data, the linkages of data, and performance of the system are optimized. Once the SLDS architecture is implemented and operational, it is important to monitor the types of queries executing in the system. This monitoring allows for tuning of the system to improve its performance, and to understand how the security model is enforcing the rules. It is not uncommon that rules will need to be tweaked as data governance continues to evolve.

Although the Virginia SLDS' stakeholders have a strong understanding of the system's baseline data governance, a number of factors will influence the need for upfront and ongoing efforts. The Virginia SLDS will add future sources and this will require both a rigorous upfront effort to minimize rework and redesign and will necessitate ongoing efforts when changes are made to the existing data sources or when new data sources are added.

*(Army Suicide Mitigation, DLA, North Dakota SLDS, Texas Education Agency, Washington Research and Data Center)*

### **3.4 Leveraging Existing Systems**

Some SLDS projects were able to leverage existing systems to become the foundation of the SLDS system; this saved time and resources during the design and implementation stages of the project.

Indiana was able to leverage an existing system, Learning Connection, and expand on its capabilities. Initially, Indiana had not planned to expand its Learning Connection portal, as it was built primarily for teacher networking and was not intended to be a working data system for other stakeholders. However, due to political conflicts, Learning Connection evolved into such a system. Indiana's LDS project was begun by the previous administration, and the state's new leadership originally planned to eliminate Learning Connection. The Indiana Department of Education, however, argued that starting over with a new data system for K-12 would not be cost-effective. In the end, Learning Connection was modified to be used as a collaboration site and as a K-12 data management system.

North Dakota also was able to leverage an existing data warehouses to avoid "reinventing the wheel."<sup>5</sup> After surveying what systems existed in the state, the North Dakota SLDS team discovered that their legacy K-12 system had the technical capabilities to form the LDS foundation. This K-12 warehouse will be expanded into an LDS and will collect information from other agencies. By building out the K-12 data warehouse into an LDS, North Dakota's team saved time and money in the project, which will enable them to focus on other technical and non-technical issues (such as linkages between other data systems).

*(Indiana Department of Education, North Dakota Department of Public Instruction)*

### **3.5 Use of Commercial Systems**

Our interviews revealed both positive and negative consequences of using commercial off-the-shelf solutions; these commercial solutions can be a benefit, saving agencies time from building its own solutions, but they can also limit the versatility and expandability of the system.

---

<sup>5</sup> Korsmo, T. (2010, October 26). Telephone Interview with Rona Jobe.

The Indiana Department of Education began their project with an Oracle platform for their data warehouse, but eventually switched to a SQL platform. . After a year's effort, the project staff realized that Oracle was not meeting their needs, was overly complicated, was not user-friendly, and was extremely expensive. The team restarted with a new solution and had to perform rework because of the commercial solution they initially chose; however, since their move to the SQL platform, the SLDS has progressed rapidly and has performed well.

The Iowa Department of Education purchased an off-the-shelf data model for their SLDS. This model was adopted prior to Iowa receiving the SLDS award, when the system was focused on the K-12 space. After deciding to expand their efforts to P-16, the SLDS team found that the data model they had purchased did not work as an effective model for the higher education data within the state. In order to integrate the higher education data, the Iowa team is investigating whether to purchase additional commercial data models or to build their own custom data models (both which will require additional financial and man hour resources).

The Texas Education Agency purchased commercial off-the-shelf solutions in order to develop a public-facing Web site where users can access the data from the SLDS. The Texas SLDS team found that commercial software provided adequate tools that allowed them to maintain the Web site, while minimizing maintenance resources.

There are few off-the-shelf solutions that are able to perform data integration on-the-fly in federated databases, but the space continues to grow. Major database vendors, e.g. IBM, Oracle, have federated database management systems that are able to assist with the integration requirement.

*(Indiana Department of Education, Iowa Department of Education, Texas Education Agency, Ariyachandra, Haas, Ramesh)*

### **3.6 Multiple Hash Keys**

Encryption of personal identifiable information (PII) using one-way hashing was discussed as a method for protecting an individual's identity. Subject Matter Experts mentioned that using various data to create hash keys can provide a number of options for greater record matching. Techniques can include combining multiple values into a single hash, or creating multiple hashes that can be used for comparison.

*(Carney, Carter, Dobler, Kleinman)*

### **3.7 Requirements and System Architecture**

Reporting and usage requirements should determine the type of architecture to be built and identifying these elements of the system early in the design and development phases will save time and money. Data warehousing specialists at Claraview,<sup>6</sup> emphasized during an interview that knowing how the system should perform and what functions will be required will drive the architecture of the system. In essence, the architecture of an LDS should be determined largely by what an agency wants it to do. The Claraview team cautioned that, as they have witnessed with other state departments of education, failure to identify and address system and stakeholder needs adequately will result in a failed or less than optimal LDS.

---

<sup>6</sup> Claraview is a business intelligence and data warehousing consulting organization. See <http://www.claraview.com/dnn/>

The Indiana Department of Education team concurred that design should be dependent upon functions, or how the agency plans to use the system. They reiterated that since all states have differing reporting requirements and needs, no one design solution will be appropriate for all. This requirement should be closely aligned with identifying stakeholder needs.

*(Indiana Department of Education)*

### **3.8 Clearly Defined Security Policies**

Instituting security policies for the protection of data to prevent the possible identification of a person is critical to the success of the system. SMEs stated that security policies and measures need to be defined clearly. Security policies, in combination with the database security, can maximize the protection of sensitive data. To rely only on database security tools would be short-sighted. It is important to review all aspects of security, including operating system hardening practices and network device configurations.

*(Dwork, Kleinman)*

## 4 Architecture Best Practice Case Studies

The goal of the Architectural Design and Best Practices Project was to provide the VLDS team with an up-to-date and relevant assessment of the best practices related to the design, development, deployment, and operation of a Statewide Longitudinal Data System.

The Center for Innovative Technology (CIT) was commissioned by VDOE to conduct research on similar longitudinal database development efforts or large data integrations across disparate organizations. Additionally, CIT was tasked to produce best practice recommendations that would include the identification of risks and impediments in building an LDS. Based on the information collected from the nine individual case studies, the project team consolidated themes, lessons learned and best practices.

### 4.1 Indiana Department of Education

State/Agency: Indiana Department of Education  
 Web Site: <http://www.doe.in.gov/data/>  
 Address: 151 West Ohio Street  
 Indianapolis, Indiana 46204  
 POC: Molly Chamberlin  
 Director of Data Analysis Collection and Reporting  
 POC Phone: 317-234-6849  
 POC Email: [mchamber@doe.in.gov](mailto:mchamber@doe.in.gov)

Case Profile
<b>Student Enrollment:</b> 1,046,147 <sup>7</sup>
<b>Teachers:</b> 62,668 <sup>8</sup>
<b>LDS Grant:</b> \$5,188,260 <sup>9</sup>

#### Background

In 2007, the Indiana Department of Education (IDOE) was awarded approximately \$5.2 million to create a comprehensive P-20 data system. For its LDS, IDOE envisioned a system that would “allow data integration at all levels and would enable stakeholders to track and to analyze student achievement and attainment from early childhood through higher education and beyond.”<sup>10</sup> The main objectives of Indiana’s LDS were to improve data quality; provide

<sup>7</sup> *State educational data profiles. (n.d.).* Retrieved from <http://nces.ed.gov/programs/stateprofiles/sresult.asp?mode=short&s1=18>

<sup>8</sup> Ibid.

<sup>9</sup> *Statewide longitudinal data system grant program - grantee state - Indiana. (n.d.).* Retrieved from <http://nces.ed.gov/programs/slds/state.asp?stateabbr=IN>

<sup>10</sup> *Indiana P-20 Comprehensive Data System. (n.d.).* Retrieved from <http://nces.ed.gov/programs/slds/pdf/Indianaabstract.pdf>

longitudinally linked data to be used to drive policy decisions; and to make the data user-friendly for teachers, principals, superintendents, and other stakeholders.

The project involved many different stakeholders, primarily: the Local Education Agencies, IDOE data warehouse, internal IDOE staff, Department of Workforce Development, and higher education (e.g., Ivy Tech, Indiana’s statewide community college network). Indirectly, the project’s stakeholders and consumers were policy makers, legislators, parents and students.

Initially, IDOE held monthly meetings for the LDS’ stakeholders. These stakeholders included representatives from every major division in the IDOE, special education, language minority, Title I, Curriculum and Instruction, Data Reporting, Student Services, Technology, 40 fellows from different school systems across the state, state administrators, etc. Because of the large size of the group, the meetings became “too involved and unproductive.”<sup>11</sup> An outside evaluator suggested performing a series of interviews with the individual groups instead of holding large stakeholder meetings.

IDOE conducted interviews with each of the stakeholder groups and asked about their vision in an LDS in terms of functionality and design. During this process, IDOE acted as intermediary and a champion for each group. As a result, not only was IDOE able to gather pertinent input from each stakeholder group, but they obtained buy-in from the stakeholders. IDOE synthesized the feedback data and created a small functional committee that assisted in the day-to-day decisions of building and designing the LDS.

### Key Take-away



**Input from stakeholders is essential to designing a longitudinal data system. The requirements set forth by the stakeholders help determine the architecture and functionality of the final system.**

The IDOE team faced a number of obstacles while building Learning Connections and the data warehouse.<sup>12</sup> First, the team found that their original platform, Oracle, was expensive and time-consuming to learn. Once the platform was switched to SQL, however, the project progressed smoothly. Another difficulty was the change in Indiana’s administration. Once the new administration took office, the LDS team was forced to defend the need for an LDS and present an overview of what a data warehouse is, how it should function; and what had been done until that point. In spite of IDOE’s presentations, the new administration was still uncertain on what to do with an LDS system, particularly Learning Connection.<sup>13</sup> After several discussions, the IDOE LDS team persuaded the administration to expand Learning Connection to become an LDS tool. The new administration responded and requested additional changes – with the evaluation

<sup>11</sup> Chamberlin, M. (2010, October 25). Telephone Interview with Rona Jobe, CIT.

<sup>12</sup> Learning Connections was built in approximately 18 months, while the warehouse was built in 15 months.

<sup>13</sup> The original Learning Connection was an interactive, networking site, and the new administration wanted to close it because they did not see its value.

process and switching from Oracle to SQL. After negotiations, the LDS team was able to retain their evaluation system, but replaced their Oracle platform with SQL. After securing the new administration's consent and approval, along with the agreed-upon changes, the linkage of the two systems, Learning Connection and the warehouse, was relatively short and seamless. Additionally, IDOE also created a help desk for Learning Connection and its public reporting system, DOE Compass; both help desks have email addresses to which users can submit questions.

### *System Design and Architecture*

Indiana's LDS is an amalgamation of multiple systems. It is comprised of three main data warehouses and portals: Learning Connection, IWIS (Indiana Workforce Intelligence System), and the IDOE Data Warehouse. Additionally, a public site, IDOE Compass shares aggregate reports and data. The IDOE Compass system accesses copied tables and rolled-up data from the IDOE Warehouse, which are shared with the public. Available data sets include number of students and teachers in Indiana as a whole, and in certain districts. Access to certain data sets is restricted to authorized users.

#### *IDOE Data Warehouse*

The IDOE Data Warehouse is an internal enterprise data warehouse that is built on an SQL platform. The warehouse project commenced two years ago and employed an Oracle platform and Oracle Business Intelligence tools. This architecture was chosen because of the recommendations of the Indiana Office of Technology, prior to the new administration. After a year, IDOE reevaluated the system's performance and concluded that the Oracle tools did not meet their needs. According to the LDS staff, the Oracle tools were "not user friendly," were "overly complicated [and] extremely expensive." Therefore, the team investigated other solutions. One product they considered as a reporting platform was SharePoint, but found the software too expensive.

When the new administration took office, the Data Analysis, Collection, and Reporting Office briefed the new CIO on the problems they had encountered in building the data warehouse and, under his guidance, the data warehouse was moved to the SQL platform. After the conversion to the SQL platform, IDOE was able to create data marts in-house. They currently are using Microsoft SSRS and SSAS and, thus far, have not encountered the problems they experienced with Oracle and noted that these tools are "very, very easy to use."<sup>14</sup>

#### *Learning Connection*

Originally, Learning Connection was built as an interactive site for Indiana teachers to exchange information on lesson plans, techniques, and resources – similar to a social-networking site. Throughout the course of Indiana's LDS project, Learning Connection evolved into a learning management tool that provides data to stakeholders at the local level. In essence, Learning Connection is a portal where teachers and administrators can access standards-based activities, share lesson plans, and communicate with other teachers. Additionally, Learning Connection also allows teachers to access their current students' longitudinal data. Currently, the system can run simple reports, but the IDOE team is working on expanding it to have more complex-reporting capabilities. Learning Connection also interacts with the Data Warehouse by pulling "copied"

---

<sup>14</sup> Ibid.

data from the warehouse. Once the data has been cleaned, checked, verified, and loaded into the warehouse, Learning Connection retrieves reports directly from the warehouse.

### *IWIS*

The Indiana Workforce Intelligence System (IWIS) is a separate data warehouse that is linked to K-12 and post-secondary data. According to the Indiana Workforce Development website, IWIS “began by integrating disparate data sets from within the Department of Workforce Development to then integrating [the] resulting new data with information from the Commission for Higher Education.”<sup>15</sup> This data warehouse is outside of the IDOE and is run by Indiana’s Department of Workforce Development. IWIS was added to the LDS project after Learning Connection and the IDOE Data Warehouse. Originally, the Indiana SLDS team planned to populate the IDOE Data Warehouse with linked data. However, after discovering that the Department of Workforce Development and members of the higher education community had systems of their own, the SLDS team decided to link them with Indiana’s LDS.

Theoretically IWEIS also will pull its data from the data warehouse; however, IDOE currently is struggling with acquiring access to data from its Department of Workforce Development, although Ms. Chamberlain did not elaborate on this point.

### *Security*

Information within the Data Warehouse is identifiable, however, when other systems pull information from the warehouse, the warehouse creates a set of tables from the identified data that is de-identified and aggregated. In essence, systems do not actually access the source data directly. For example, Learning Connection only accesses tables that have been created and copied from the warehouse. Additionally, the warehouse and Learning Connection utilizes role-based permissions; educators in Learning Connection have access only to their current students and administrators have access only to the students currently attending their schools.

Moreover, the source data in the warehouse can be accessed only by certain IDOE personnel who have appropriate permissions – approximately four people. Ms. Chamberlain declined to disclose what other security measures have been implemented.

### *Data Usage and Reporting*

The warehouse houses five years of data, which represents approximately one million public school students’ records and 65,000 non-public school students’ records. Assessments data are generated once a year – this includes enrollment and other data required to generate the state report card and reports to the federal government.

The public may view prepared aggregated data sets by school and by district, as well as public reports through the Compass data site. The system accesses copied tables (rolled up data) to generate these reports and aggregated data.<sup>16</sup> Certain data from the IDOE Compass site are accessible only to registered users. However, other unidentified, aggregate data sets that are not readily available through the site can be requested. Depending upon the size of the request, this

---

<sup>15</sup> Chamberlin, M. (2010, October 25). Telephone Interview with Rona Jobe, CIT.

<sup>16</sup> The system is not accessing de-identified data, but rather aggregated data.

data usually can be provided within 2 to 14 business days.

(<http://compass.doe.in.gov/Dashboard.aspx?view=STATE&val=0&desc=STATE>)

Researchers also may submit requests for large data sets. IDOE has a number of “canned reports” for researchers (e.g., enrollment by school in the last five years). However, if researchers request student-level data that is de-identified, this request is processed by the legal department and must sign a data sharing agreement. Once the legal department approves the researcher’s request, s/he may use the online data request system for IDOE Compass. The request is entered into a queue and IDOE personnel retrieve and review the de-identified data before it is released. Currently, the IDOE team is expanding the system’s reporting capabilities. They note that had they taken into account the types of reports and departmental requirements from the LDS in the beginning, building and expanding the system would have been easier.

### *Lessons Learned*

Throughout Indiana’s LDS project, the LDS team found practices that have helped along the way. First, IDOE discovered the efficacy of building upon legacy systems like Learning Connection because it saved time and money. IDOE also found that prohibiting other linked systems to access source data enhances security and preserves a consistent “true” record. Data cleaning is imperative. Lastly, achieving stakeholder buy-in and gaining feedback is important in building a system. Interviewing different stakeholders and representatives individually provided IDOE substantive information on stakeholder requirements (e.g., types of reports and how the system should perform). Furthermore, IDOE found that gaining stakeholder buy-in is also important for a smooth LDS implementation.

#### **Key Take-away**



**Building on legacy systems saves time and money, e.g., turning Learning Connection, initially a teacher-networking system, into a data management tool rather than eliminating the system and starting over.**

## 4.2 Iowa Department of Education

State/Agency: Iowa Department of Education

Web Site:

[http://www.iowa.gov/educate/index.php?option=com\\_content&view=article&id=1691:edinsight&catid=45:data-collections&Itemid=2490](http://www.iowa.gov/educate/index.php?option=com_content&view=article&id=1691:edinsight&catid=45:data-collections&Itemid=2490)

Address: 400 E 14th St  
Des Moines, Iowa 50319

POC: Jay Pennington  
Bureau Chief

POC Phone: 515-281-4837

POC Email: [jay.pennington@iowa.gov](mailto:jay.pennington@iowa.gov)

### Case Profile

**Student Enrollment:** 487,559<sup>17</sup>

**Teachers:** 35,961<sup>18</sup>

**LDS Grant:** \$8,777,459<sup>19</sup>

### Background

In 2008, Iowa initiated a project to create EdInsight, the Iowa Department of Education's (IDE) K-12 centralized data warehouse. EdInsight integrated seven years of historical data from Project EASIER (student level enrollment and curriculum data), IMS (special education data), and the Iowa Testing Program (student assessment data). The initial budget for the project was \$1.2 million and had a total implementation cost of \$2.9 million through FY2009. In May 2009, the project was funded by an \$8.78 million SLDS grant which would be used to increase the scope and functionality of EdInsight to be interoperable with postsecondary data systems or to create a consolidated P-16 data system. The LDS team plans to add additional sources of information such as teacher, financial, transcript, workforce, disaster mitigation, and additional assessment data. EdInsight is still in its statewide rollout phase.

### System Design and Architecture

IDE decided to use a commercial off-the-shelf (COTS) data model for the EdInsight project because this particular data model was designed, specifically, for use in the K-12 space. During the design process, the IDE team discovered that some of the data within the system did not fit the COTS model, particularly, the post-secondary data. Eventually, however, this post-secondary data was integrated into EdInsight.

<sup>17</sup> *State educational data profiles.* (n.d.). Retrieved from <http://nces.ed.gov/programs/stateprofiles/sresult.asp?mode=short&s1=19>

<sup>18</sup> Ibid.

<sup>19</sup> *Statewide longitudinal data system grant program - grantee state - Iowa.* (n.d.). Retrieved from <http://nces.ed.gov/programs/slids/state.asp?stateabbr=IA>

### *Security*

Security is managed through the use of role-based access and training.

### *Data Usage and Reporting*

EdInsight's data is used to conduct analyses and produce reports for education stakeholders, such as the IDE staff who are granted access to data in preformatted reports and advanced data analyses, depending on their role and permissions. Currently, over 150 users have been trained and given access to the system and more than a dozen pre-formatted reports have been developed. At this time, there are no plans to allow users to perform ad-hoc queries.

A portion of the SLDS grant will fund the creation of a public portal. This portal will provide aggregate-level data that will be accessible through the Web; however, the portal has not yet been developed.

### *Lessons Learned*

Mr. Pennington stated that gaining buy-in at the regional level was critical to the current success of the project and will continue to play a key factor during its statewide rollout. He observed that COTS product did not meet IDE's needs and took longer to load and to format the data. To combat this problem, the team is investigating whether to purchase or develop additional modules that will fit the post-secondary and workforce data that IDE intends to integrate into EdInsight.

#### **Key Take-away**



**Commercial off-the-shelf solutions must be evaluated carefully and set against the system's current and future requirements in order to attain their cost- and time-saving benefits**

### 4.3 Army Suicide Mitigation Project

State/Agency: Data Strategies  
 Web Site: <http://www.datastrategiesinc.com>  
 Address: P.O. Box 772  
 Midlothian, Virginia 23113  
 POC: Susan Carter  
 Managing Partner  
 Kevin Corbett  
 Managing Partner  
 POC Phone: 804-965-0003  
 POC Email: [SCarter@DataStrategiesInc.com](mailto:SCarter@DataStrategiesInc.com)  
[KCorbett@DataStrategiesInc.com](mailto:KCorbett@DataStrategiesInc.com)

Case Profile
<p><b># of Records:</b> Unavailable</p> <p><b>Project Budget:</b> Unavailable</p>

#### *Background*

Due to an increase in suicides, the United States Army hired Data Strategies to design and pilot a prototype data system that would gather information from disparate sources in order to identify predictors of potential suicides. The goal of the pilot project was to utilize this information to establish a means to stem the number of suicides and suicide attempts. In order to achieve this goal, Army leaders realized that they would need an integrated data environment that would provide accurate and reliable data for analysis.

The project team's challenge was to develop a system that would rely on numerous databases,<sup>20</sup> both government and private, that had not been linked previously. The system first must analyze historical data of suicide cases from 2001 to 2008 in order to determine if there are any commonalities. These commonalities then will be matched against the records of current soldiers in the hope of identifying those who may be at a high risk for suicide.

In terms of database management systems, although the sample group was relatively small, the size of the records was large. Because of the sensitivity of the topic and the need to ensure the soldiers' privacy and the security of their information, personal, identifiable information was removed or de-identified. Further, since the Army had no stringent performance requirements, such as ad-hoc queries into the system, the majority of the analyses were performed on historical

<sup>20</sup> Some of the data sources included are Army, financial and medical. The project team must negotiate HIPAA requirements, which will impact the Army's ability to aggregate the data.

data that was static. As a result, the time from query to data delivery could be allowed to take days.

### *System Design and Architecture*

The final design of the system had not yet been determined at the time of the interview, partly due to the fact that many of the leaders of the planned data sources had not yet signed memoranda of understanding (MOU) or data sharing agreements. Because these data sources were from different industries, they did not follow a shared schema, governance or, in many cases, data types. At that time, Data Strategies planned to investigate a number of design types that would allow the Army a choice of data types in the future. The project team considered various architectures and schemas, while remaining open to various data types (e.g., Excel, Oracle, and flat file types) to ensure that the system could be flexible and expandable.

A sandbox environment was created as a centralized data warehouse that copied data from its data sources. This warehouse was used by Data Strategies because they were not allowed direct access to the data sources for the prototype development, however, the sandbox allowed Data Strategies to mimic systems that could be centralized, distributed or federated database management systems. For the purpose of the prototype, queries were not submitted live across the internet but, instead, used the sandbox environment. This meant that although real performance of the system was not measured, this was acceptable since speed performance was not a requirement of the system at this stage of development.

### *Security*

In order to meet the security requirements set forth by the program, the Data Strategies team de-identified the records from the various databases, but still had to be able to link the data to perform the analyses. To accomplish this, the team created unique identification (ID) numbers. They were able to link this unique ID to the records of each of the databases in the following two ways:

1. They searched for an individual's records that contained an existing unique ID and then pushed that unique ID to all the remaining data sources.
2. They found a unique ID contained within each of the data sources and created a table of those IDs at the central data warehouse.

Due to the relatively small subject group, both of these solutions worked.

### *Data Usage and Reporting*

The goal for the system is to have de-identified aggregate data that will allow only authorized users within the Army to analyze the data. The data and reports were not made available to the public or to any participating data sources.

### *Lessons Learned*

Ms. Carter explained that although there were many architectural and technological barriers to the project, the single most complex obstacle to overcome within this project was the management among the various data sources. Although the various Army agencies were under a mandate by the Secretary of the Army to participate in this pilot program, the external agencies

providing information were not. The MOU and data sharing agreements<sup>21</sup> had yet to be negotiated and signed and it was necessary that these documents be executed prior to Data Strategies accessing the data source information and integrating it into the army suicide system.

Due to the number of data sources and the underestimation of resources needed to manage stakeholders and execute these tasks, many of the MOUs and data sharing agreements were not signed during the pilot project. A final implementation of this system would require MOUs and data sharing agreements that could take years to be signed. Ms. Carter recommended that organizations planning to construct a longitudinal database make sure that they plan to commit resources to the development of the MOUs and data sharing agreements as well as the management of the various stakeholders well in advance of the project launch.

### Key Take-away



**Data governance and stakeholder management are upfront efforts but they also require ongoing efforts that should not be overlooked. These two efforts are essential to ensure the reliability and expandability of the system**

---

<sup>21</sup> These agreements determined who would participate, what data would be shared, how it was to be utilized, and where it could be stored.

## 4.4 Texas Education Agency

State/Agency: Texas Education Agency  
 Web Site: <http://www.texaseducationinfo.org/tpeir/>  
 Address: Information Analysis, TPEIR Group  
 Texas Education Agency  
 1701 North Congress Avenue  
 Austin, Texas 78701  
 POC: Brian Rawson  
 Director, Statewide Data Initiatives  
 Nina Taylor  
 Director of Information Analysis  
 POC Phone: 512-463-9437  
 512-475-2085  
 POC Email: [Brian.Rawson@tea.state.tx.us](mailto:Brian.Rawson@tea.state.tx.us)  
[Nina.Taylor@tea.state.tx.us](mailto:Nina.Taylor@tea.state.tx.us)

Case Profile
<b>Student Enrollment:</b> 4,752,148 <sup>22</sup>
<b>Teachers:</b> 327,905 <sup>23</sup>
<b>LDS Grant:</b> \$18,195,078 <sup>24</sup>

### Background

In 2001, the Texas Legislature funded a project that would build an integrated data repository for the Texas Education Agency (TEA), the Texas Higher Education Coordinating Board (THECB), and State Board for Educator Certification (SBEC). The project became known as the Texas PK-16 Public Education Information Resource (TPEIR) Project. Half of TPEIR's original \$7 million appropriation for the public access initiative was set aside for FY2002 and FY2003.<sup>25</sup> The system that resulted from the project team's work integrates the data from disparate data sources at each of the participating agencies. These data include student, educator, and organizational data from as far back as 1989.

TPEIR was designed to ensure that stakeholders within Texas would have access to high quality data using an efficient and effective method to obtain it and would link student data from early childhood through postgraduate study to allow for longitudinal analysis that would identify patterns and trends within the Texas public education system. Data from TPEIR was planned to be available to internal staff as well as to the public.

<sup>22</sup> *State educational data profiles*. (n.d.). Retrieved from <http://nces.ed.gov/programs/stateprofiles/sresult.asp?mode=short&s1=48>

<sup>23</sup> Ibid.

<sup>24</sup> *Statewide longitudinal data system grant program - grantee state - Texas*. (n.d.). Retrieved from <http://nces.ed.gov/programs/slds/state.asp?stateabbr=TX>

<sup>25</sup> The final total cost of the project was \$6.1 million, with \$1.75 million spent in FY2002, and \$4.35 million spent in FY2003. In May 2010, Texas won an \$18.2 million SLDS grant, the second highest grant amount awarded.

Project management of TPEIR was a complex and ongoing effort. The management of the system was handled by two advisory groups, the Interagency Steering Committee (ISC) comprised of the Information Resources Managers of each agency and the Technical Advisory Group (TAG) comprised of the project managers of each agency. The ISC met twice a month to determine policy, review risks, and resolve issues, and the TAG met weekly to determine the technical infrastructure, plan the practical implementation, and resolve technical issues.

### *System Design and Architecture*

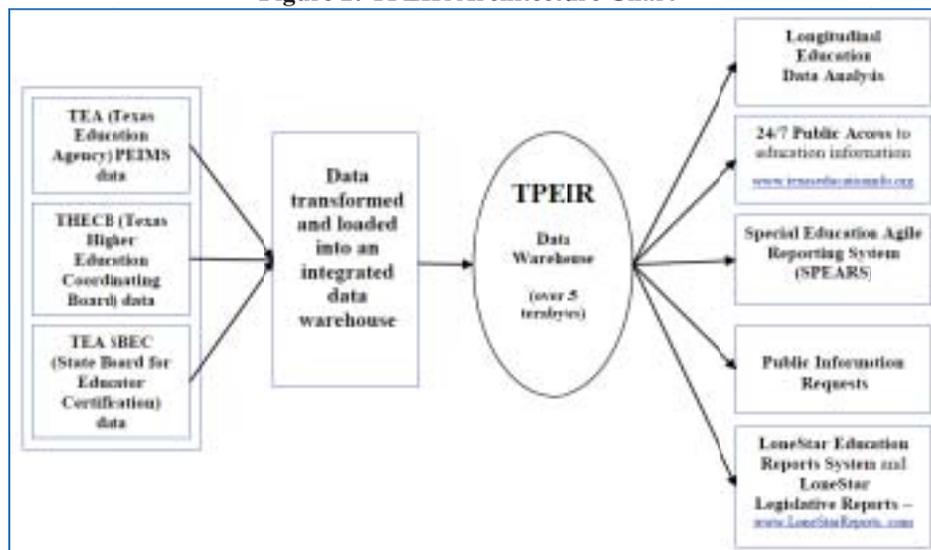
TPEIR was designed with two distinct data repositories. One repository housed aggregated data<sup>26</sup> that was de-identified and approved for public release. In order to comply with federal and state standards, the other repository contains confidential, student-level education data that is available only to authorized users.

The actual development of the system was outsourced to an outside vendor. The resulting custom system design adopted a combination of the Ralph Kimball (i.e., a conglomerate of data marts) and Bill Inmon methodologies (i.e., a single data warehouse) as the foundation for the data warehouse, which was similar to that of the TEA K-12 data warehouse.

The data warehouse stores facts/metrics within fact tables and codes within dimension tables. An AIX server was used during the development and testing processes, but the final data collection was moved to a production server.

TPEIR currently integrates data from two data sources into a centralized database, but its architecture framework allows for new data sources to be added in order to enhance the power of the system. Figure 2 illustrates the TPEIR architecture framework.

**Figure 2: TPEIR Architecture Chart<sup>27</sup>**



<sup>26</sup> This data can be accessed at <http://www.texaseducationinfo.org>.

<sup>27</sup> Texas Education Agency, Information Analysis Division. (2010). *Texas PK-16 public education information resource* Retrieved from [http://www.texaseducationinfo.org/tpeir/TPEIR\\_Documentation.pdf](http://www.texaseducationinfo.org/tpeir/TPEIR_Documentation.pdf)

### *Security*

The team ensured the system's security by creating dimension tables that used surrogate keys that were arbitrary, system generated values as unique identifiers. These keys were used to perform the linkages in the system.

### *Data Usage and Reporting*

The system's report component uses Crystal Reports solution. The TPEIR data is available to the public and to authorized stakeholders. The publicly available data are used to report on Texas public high school graduation; Texas college and university admissions, enrollment, and graduation; teacher certification, employment, and retention; and school district employment. A complete list of publicly available reports can be found at: <http://www.texaseducationinfo.org/> Authorized TEA staff members use Rapid SQL or SAS to run queries against the data and generate reports or extract data to be stored in files. These results of these queries are returned as quickly as a few seconds while larger queries may take several minutes.

### *Lessons Learned*

The integration of the data from three different agencies and the conduct of multiple data collections (while preserving the original data) was a problem that the TPEIR team faced early in the planning of the SLDS. It was important to preserve the original data so that each agency could recreate historical results, if necessary. The team conformed data across the agencies and defined standards that would applied to current and future data collections. They maintained regular meetings of the Interagency Steering Committee (ISC) and Technical Advisory Group (TAG) to exchange information, review changes, resolve issues, and establish consensus.

#### **Key Take-away**



**For systems that require the source data to maintain its data integrity, the data governance is critical. It is through the stringent standard and rules definitions that the data sources are able to share their data for use in the SLDS while preserving the original database's system.**

Another problem the project team faced was the implementation of the public-facing Web site that would allow the public to access data. TEA wanted to minimize the maintenance requirements for this Web site. To accomplish this, the TPEIR team utilized commercial off-the shelf software and minimized the customization of software tools to maintain the Web site. These tools allowed the developers to utilize metadata, common educational terminology, online help pages, standard reporting formats, simple navigation, and alternatives to view the data in text and/or graphic formats without large expenditures.

As a result of following these best practices, the TPEIR's final expenditures were nearly twelve percent under budget.

## 4.5 DLA Data Convergence and Quality Project

State/Agency: Data Strategies  
Web Site: <http://www.datastrategiesinc.com>  
Address: P.O. Box 772  
Midlothian, Virginia 23113  
POC: Susan Carter  
POC Phone: 804-965-0003  
POC Email: [SCarter@DataStrategiesInc.com](mailto:SCarter@DataStrategiesInc.com)

### Case Profile

**# of Records:** 7 million base records (each record had approximately 15-20 associated records)

**Project Cost:** \$2.5 million over 5 years

### *Background*

In 2002, the Defense Logistics Agency (DLA) hired Data Strategies to vet its process and systems issues in implementing a Business Systems Modernization (BSM) program. This five-year project was part of a larger system overhaul that DLA implemented during a period of over 10 years. The BSM program was implemented to upgrade the procurement and financial systems that managed DLA's supply chain management processes. The new process required DLA to deliver accurate information and data for business, profiling standards, business rules, and processes. However, because the three centers have evolved over time, it was difficult to merge them.

Originally, DLA began with three main supply “centers” that performed the same functions on different items. When the project first began 50 years ago, the three centers had identical architecture and business processes. Over time, however, the three centers evolved and began using different methods and business rules. In 2002, DLA initiated a massive merger of the three centers. The goal was to make the centers interoperable and compliant with the new business rules DLA was developing in order to have centralized. In other words, although the data centers were physically separated, the data was, in a virtual perspective, to be integrated and located in one central place, since users needed the ability to retrieve procurements that were located in more than one data center.

This integration meant that the data had to appear to the user to be in one place so that they could query across the centers. The project had three main stakeholder groups, which are the “owners” of each data centers, and a fourth entity, an umbrella record center called the Logistics Information Group (LIG). Logistics Information Group approved everything that was done to the system and records, as well as any cleansing. Moreover, this entity maintained records of what items the Defense Department could purchase and their negotiated prices. In essence, Logistics Information was the gate-keeper of what items could be procured. The LIG information became known as the “golden record.” Although this group owned the “golden record,” they were

not users of the system. While the three supply centers had their own functions and governing teams, all data had to be compliant and match with the “golden record.”

### *System Design and Architecture*

It became clear that DLA needed to develop a set of uniform business rules that would govern all centers. There was a tremendous amount of resistance from each of the centers as each group was unwilling to sacrifice control or autonomy. To eliminate this resistance, Data Strategies facilitated discussions among the centers in order to reach a consensus in creating the business rules. Data Strategies worked with each of the disparate data source owners to understand not only how each of the centers and systems captured, processed, and stored data, but also how they would each need to interact once all systems were integrated.

The process involved negotiations between the DLA headquarters and the centers on how the data is used, how it should appear, and what specific exceptions in the rules that would be required because of each center’s unique item/record types and security requirements. In the end, because of the definitions provided by DLA and the feedback from the centers, Data Strategies recommended a single set of business rules, with defined exceptions in each center.

Once the business rules were created, Data Strategies surveyed each data center and assessed its level of data cleanliness. The assessment was based on the new set of business rules. During the data cleansing process, the team:

*“...ensured that the source data and the converged data maintained their independence without losing either context. Data Strategies staff analyzed the business rules associated with the data; identified the rules and metrics required to validate that data would meet the new business rules; and created automated routines to run complex queries that analyzed and identified records that had anomalies. The results were displayed in both summary and detailed reports that showed the anomalies and the recommended solutions. The solution also included an approach using extensive heuristics and pattern matching code to overcome embedded data issues.”<sup>28</sup>*

The base population of data was 7 million and each one had 15 to 20 associated records. Data Strategies divided them into lots and evaluated each lot every 2 weeks over 18 months. This did not reduce the quantity of data, but ensured that it was compliant and collaborative.

The three data centers were integrated in virtual space so that users might access and query data and reports, regardless of the source. For the ultimate end-user, Data Strategies created narrow, role-based views of the personal systems so that users could view the data that pertained to them and for which they were cleared. The reports comprised of HQ-level statistics and trends that outlined the current status of the data quality as a whole (outlining the risk to the migration success) provided detailed information at the Source Owner, Table, and Attribute levels to identify where the largest issues were.

---

<sup>28</sup> Data Strategies. (n.d.). “DLA Data Convergence and Quality Project.”

Ms. Carter credits much of the project’s success to the solution’s capability to create detailed and dashboard views of the results as well as its recommendations for problem resolution.

### *Lessons Learned*

For large data integration projects, Ms. Carter recommends several best practices.

Ms. Carter’ first recommendation is to mentor the stakeholders involved in the data integration, so that they understand the history of what is being done and why it is being done. This gives these groups information they need to make decisions regarding the project. Conversely, project implementers and leaders must solicit feedback from stakeholders in order to understand stakeholder requirements as owners and users of the data. This two-way communication will help an organization to establish effective data governance rules.

For Data Strategies, informing the three center’s stakeholders of the project’s purposes and receiving feedback gave them an advantage in bringing the three centers together. As a neutral intermediary, Data Strategies was well received because stakeholders felt they had a voice in the development of the new business rules. Ms. Carter explained that, “A neutral approach is important in overcoming political silos. In the end, if the different agencies’ needs are not met, the new system will be useless.”<sup>29</sup>

Throughout a project, there must be good communication between parties that is conducive to a collaborative environment. A constant review of the project goals and status is essential to ensuring the team is on track. The way the architecture is established in the beginning is important because it lays out the foundation for the rest of the project. It is important for the system designers to understand the purposes of the warehouse before they map and design the architecture. Moreover, a complex project necessitates the engagement of an open vendor that will not constrain the design of the data architecture or the project. Finally, VDOE must ensure the technology to be used can be used by everyone and must keep the technology simple for longevity.

#### Key Take-away



**Managing stakeholders – gathering input from them and being their advocate is imperative in building business rules. Having input from stakeholders and acting as their “champion” results in less friction among stakeholders.**

<sup>29</sup> Carter, S. (2010, November 5) Telephone Interview with Rona Jobe, CIT.

## 4.6 NORC at the University of Chicago Data Enclave

State/Agency: NORC at the University of Chicago Data Enclave  
 Web Site: <http://www.norc.org/DataEnclave/>  
 Address: 1155 East 60<sup>th</sup> Street  
 Chicago, Illinois 60637  
 POC: Timothy Mulcahy  
 POC Phone: 301-634-9330  
 POC Email: [mulcahy-tim@norc.org](mailto:mulcahy-tim@norc.org)

### Case Profile

**# of Records:** No actual number given, but at any given time, the system processes 40 million records

**Project Cost:** \$750,000 (initially)

### Background

The National Opinion Research Center (NORC) Data Enclave is a “secure virtual environment for storing and analyzing sensitive microdata.” The Enclave provides a confidential, protected environment within which authorized researchers can access sensitive micro-data remotely.

A brief summary from the NORC website:

*“While public use data can be disseminated in a variety of ways, there is a more limited range of options for disseminating sensitive micro-data that have not been fully de-identified for public use. Some data producers have sufficient economies of scale to develop advanced in-house solutions that serve the needs of external researchers, but most lack the resources to archive, curate, and disseminate the datasets they have collected. The NORC Data Enclave provides our partner organizations a secure platform where they can both host and build a research community around their data.”<sup>30</sup>*

The NORC Data Enclave<sup>31</sup> was established in the early 2000s; however, the build-up to the project can be traced to decades of history. There had been some movement within government agencies and other organizations to provide access to microdata with sensitive content to researchers and to research organizations. In 2002, the Confidential Information Protection and Statistical Efficiency Act was passed; this was a mandate to all the federal statistical agencies to develop a plan “to provide some level of access to some parts of their agencies’ microdata.” In 2006, the National Institute on Standards and Technology (NIST) released a Request for Proposal (RFP) that described the need to conceptualize and build a secure remote access

<sup>30</sup> Data Enclave – NORC at the University of Chicago. (n.d.) Retrieved from <http://www.norc.uchicago.edu/DataEnclave/>

<sup>31</sup> The Enclave’s design and development costs were approximately \$750,000. The third iteration is planned for a February 2011 and may receive an additional \$500,000 to \$750,000 in funding.

modality that could provide both on-site and remote access to microdata as well as direct access to the raw microdata.

Originally, statisticians, lawyers, and agency leaders were very concerned about allowing researchers access to raw data and these groups developed plans to perturb the data prior to allowing researchers access. However, in 2006, there was a significant change in thinking and decades' worth of thought engineering and science on how to perturb data to be ready for researchers. The new model allowed researchers, other governmental agencies and private sector organizations to have access to the actual raw microdata as opposed to allowing them access only to perturbed data.

This shift in policy occurred when NORC proved<sup>32</sup> that there were remarkable differences in research results if a researcher used perturbed data rather than raw data. In some cases, researched based on perturbed data yielded results opposite from what they would have been had researchers been allowed to access the raw data.<sup>33</sup> This revelation caused leaders in government agencies to question whether previous policies and programs created from research performed with perturbed data were based upon false assumptions. As a result, these statisticians, lawyers and agencies who originally opposed the idea of providing raw data to researchers changed their stand on the matter.

The challenge became to find the true results and gather the data to be available in the public domain, while at the same time protecting the confidentiality of the provider of the data or survey. Shortly thereafter, the project was launched. It is sponsored by the National Institute of Standards and Technology, the Kauffman Foundation, the Department of Agriculture, the National Science Foundation, and the Annie E. Casey Foundation.

The NORC Data Enclave team's goal was to "provide a secure remote access modality that was both sophisticated, technologically and operationally, and reasonably cost and met the replication standards and ability to push the risk of breach as far down to zero as possible."<sup>34</sup> Additionally, they aimed to provide remote access. Until the NORC project, access to sensitive data for researchers was a cumbersome and time-consuming process. Researchers had to access and perform analyses on the data on site, and were not allowed to leave the building with any data. The process required that researchers be mailed their data and analysis after an internal statistician carefully reviewed their analyses. The NORC Data Enclave aimed to relieve that burden on researchers.

In summation, the aim of the new system was to share social science data in a secure manner. NORC plans to promote access to sensitive business microdata; protect confidentiality; archive, index, and curate microdata, and encourage researcher collaboration.

### *System Design and Architecture*

The Enclave is constantly being enhanced. The Enclave had a soft launch in 2006, with a 6-month incubation period (July through December 2006). Researchers and focus groups consulted in

---

<sup>32</sup> Mulcahy, T. (2010, October 28). Telephone Interview with Rona Jobe, CIT.

<sup>33</sup> Ibid.

<sup>34</sup> Mulcahy, T. (2010, October 28), Telephone Interview with Rona Jobe, CIT.

December 2006. The NORC staff collected and responded to feedback and, afterward, opened the Enclave in March 2007. In designing and building the Enclave, NORC employed a 14 to 16-person team comprised of engineers, researchers, information technologists, and metadata people, among others, and began with white-boarding and mock-ups. The initial process of white-boarding lasted many months and was the result of numerous meetings, feedback and resolutions. The design process was reiterative, although no scenarios of how the data would be used were created because of the infinite possibilities of research questions. Thus, the focus during the design process was the possibility of a convenient secure remote access system that was virtually impregnable.

The core infrastructure is, essentially, a standard implementation with CITRIX security requirements for remote access ability. CITRIX provides layers of security for remote access in general; however, the Enclave team customized the environment by adding specialized tools that researchers would need to perform certain types of analyses (e.g., statistical packages). For data management, NORC has built into the system a way of packaging the data for the researchers, tracking what researchers are doing with the data, etc.

### *Security*

Before data is loaded onto the Enclave, it is cleaned so that it is harmonized with the data sets already in the system. Every data set that enters the Enclave must go through a DDI (Data Documentation Initiative) checklist to be DDI-compliant and SDMX (Statistical Data and Metadata Exchange)-compliant for time-series data. NORC also employs a metadata services team and an IHSN microdata management toolkit.

NORC utilizes a portfolio approach to security measures by bundling multiple protections. The system uses the Citrix client's built-in security measures for the front-end security.

### *Data Usage and Reporting*

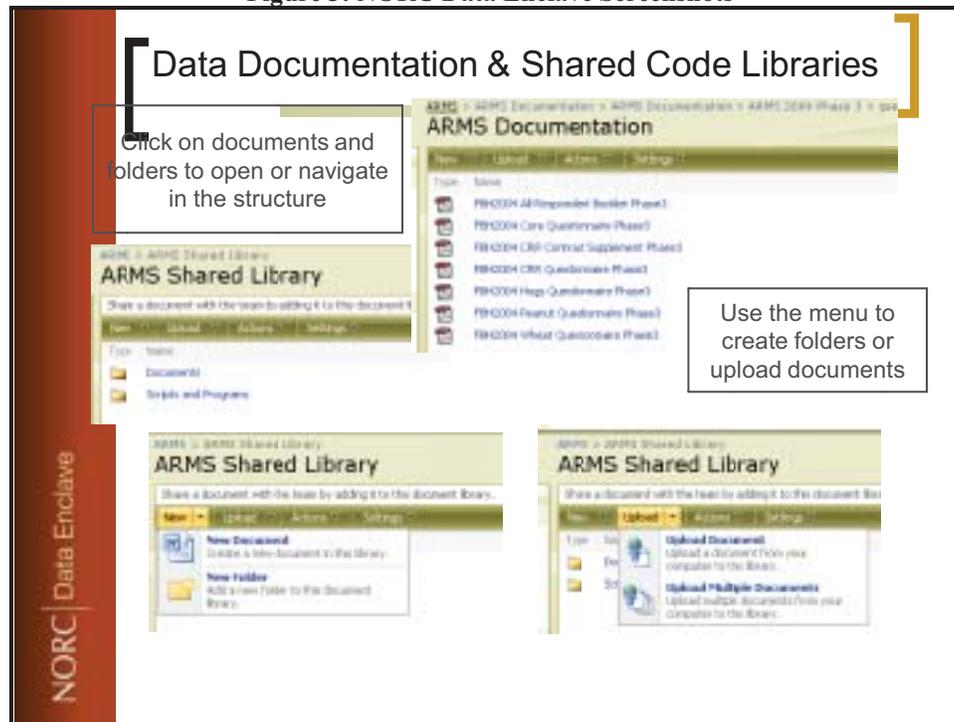
To access data in the Enclave, every researcher first must go through a vetting process by each of the sponsors. Each sponsor decides the rules on who is eligible and who will be authorized to access the data. Once the researchers get past this vetting process, they must submit proposals and substantiate why they need raw data from the Enclave – in essence, why the public use-data is not adequate for their research. For federal statistics data sets, researchers must substantiate that their research is within the mission of the federal agency and that the data required is for pure research purposes – not for marketing, law enforcement, etc. The proposals must include their planned statistical and dissemination methods and potential outlets. In general, there are several contractual steps before any researcher can be granted access to raw data. Currently, some of the data available within the Enclave include:

- NIST-TIP
  - ATP Survey of Joint Ventures (JV)
  - ATP Survey of Applicants
  - Business Reporting Survey Series (BRS)
- USDA/ERS/NASS
  - Agricultural Resource Management Survey (ARMS)
- National Science Foundation
  - Survey of Earned Doctorates (SED)

- Survey of Doctoral Recipients (SDR)
- Kauffman Foundation
  - Kauffman Firm Survey (KFS)
- Annie E Casey
  - Making Connections Survey (MC)

Access to the Enclave is gained through a Web site portal at <https://enclave.norc.org>. Users are required to download the CITRIX Client to their desktops and use the usernames and passwords that were provided to them. Inside the Enclave are collaboration tools, statistical software packages, discussion forums, etc. (See Figure 3: NORC Data Enclave Screenshots for sample screen shots from the NORC Enclave presentation). However, although researchers are able to collaborate with other researchers within the Enclave, they are not allowed to talk to one other or share data.

**Figure 3: NORC Data Enclave Screenshots**



When data is provided to researchers, certain information is stripped from it – e.g., social security numbers, addresses, birth dates, and other obvious identifiers. The level of data stripping is dependent upon which agency has supplied it. In some cases, agencies will allow access to the raw microdata, though with some data noise. The Enclave aims to provide as much data granularity to its users as possible, so that their research results are as true as possible. The system currently has more than 200 researchers across various sponsor areas. The number of users is expected to increase to more than 300 within the next six months and, potentially, to more than 600 in the next two or three years.

### *Lessons Learned*

For Virginia, one solution will not work – its LDS will require a mixture of different solutions. There are several options that the state could pursue, especially if data will be shared with the public, but security and confidentiality must be preserved. One method of sharing data with the public is through a batch execution job. In essence, a researcher will submit a research question or query and an internal staff member will run the analysis and proof for disclosure, thereafter returning the output to the researcher.

Mr. Mulcahy advises that “it is more efficient now for people to build their own” systems. Moreover, the different components in a system e.g., business intelligence tools, should complement one another and not compete with each other. When planning to build, “keep your options open.”

In the area of project management and building the architecture, ensure that staff and personnel have defined roles and responsibilities. Ensure that the project has a well-defined plan and goals. NORC notes that in such technical projects, the technical issues are the easiest to overcome; project management and stakeholder management are the more difficult tasks.

#### Key Take-away



**In building an LDS that shares information with the public, keep your options open and do not be constricted to one product solution.**

## 5 Subject Matter Expert Interviews

The research team interviewed ten Subject Matter Experts (SMEs) in order to provide the VLDS team with specific feedback on barriers, risks, design issues and opportunities associated with implementing a large data system. The following individuals were interviewed and, based on the information collected during these interviews, the project team consolidated themes into the key points presented in this section. Specifics key points from each interview are provided in Section 3.

SUBJECT MATTER EXPERT	AREA OF EXPERTISE
Dr. Bhavani Thuraisingham	Information Security Information Management Data Management, Mining, and Security Data Mining for Counter-Terrorism.
Mr. Paul Carney	Higher Education Building Internet-Based Services
Mr. James Campbell	Large Data Integration Technology and Data Flow
Ms. Susan Carter	Data Management Information Technology New Technology Research
Mr. Raj Ramesh	Information Technology Software Product Development Enterprise Database Systems/ eLearning Systems Data Warehousing Very Large Database Systems (VLDB) Web-based Collaborative eLearning Systems Enterprise Architectures, Portals and CRM
Dr. Ron Kleinman	XML Java
Mr. Peter Dobler	Software Development Sybase SQL
Dr. Laura Haas	Computer Engineering Systems Design VLDB (Very Large Database Systems)
Thilini Ariyachandra	Information Systems Business Intelligence Data Management and Modeling Impacts of Social Networking
Dr. Cynthia Dwork	Privacy Preserving Data Analysis Differential Privacy Cryptography Distributed Computing

## 5.1 *Dr. Bhavani Thuraisingham*

Title: Director, Cyber Security Center  
Professor of Computer Science  
Organization: University of Texas at Dallas  
Phone: 972-883-4738  
Email: [bhavani.thuraisingham@utdallas.edu](mailto:bhavani.thuraisingham@utdallas.edu)

### *Background*

Dr. Thuraisingham received IEEE Computer Society's prestigious 1997 Technical Achievement Award for "outstanding and innovative contributions to secure data management". Her research in information security and information management has resulted in over 60 journal articles, over 200 refereed conference papers, and three US patents. She is the author of seven books in data management, data mining and data security including one on data mining for counter-terrorism.

### Key Take-away



- Federated model works best when the domain of questions is not well known.
- It is important that the data governance be well thought out so that the access controls across the data sources is consistent.

### *Summary*

During the course of the interview, Dr. Thuraisingham made two strong points; one related to the performance of a federated model, and the other related to data governance. She stated that if the domain of questions to be answered by the system is not well known, then the best distributed database model would be federated.

Her major concern was in the area of data governance. Dr. Thuraisingham stated that it is important that the data governance be well thought out so that access controls across the data sources are consistent. She further stated that "You need to ask the question: As data moves up the hierarchy (via joins), does the governance model still work?"

## 5.2 Paul Carney

Title: Vice President, Technical Services  
Organization: Natural Insight  
Email: [pcarney@naturalinsight.com](mailto:pcarney@naturalinsight.com)

### *Background*

Combining best technical practices with management know-how, Mr. Carney has excelled in the fields of higher education, consulting and Internet-based services. He launched his first Internet business in 1997, and has since helped build additional Internet-based service organizations in both consumer and business environments. More recently, Mr. Carney oversaw the development of the Natural Insight solution, a robust resource for managing and optimizing distributed workforces.

### Key Take-away



- Consider a virtualization compute model to manage processing requirements.
- Consider creating multiple hash values based on various personal identifiable information elements.
- Beware of log files created on a system containing transactional data.

### *Summary*

Mr. Carney's experience with large scale distributed systems exposed him to the use of virtualization as a means to manage computing resources. Mr. Carney felt that when the resource utilization is unknown, virtualization should be considered to manage processing requirements for the system.

Some of his customers include financial institutions that require protection of personal identifiable information (PII). He stated that creating multiple hash values based on various personal identifiable information elements should be considered in the implementation architecture. Mr. Carney's experience demonstrated this technique allowed for multiple opportunities to find a match across the various data sources. When the discussion moved to privacy and the need to protect the individual, Mr. Carney felt that applying a hash algorithm to the PII was a big step in archiving that requirement.

He expressed concern for log files on computing systems. All log files created on a system need to be evaluated for the type of information they contain. It is possible that some log files (e.g., operating system, application) could contain transactional data that would violate privacy policies.

### 5.3 *James Campbell*

Title: Implementation Strategist  
 Organization: SIF Association  
 Phone: 202-607-5491  
 Email: [jcampbell@sifassociation.org](mailto:jcampbell@sifassociation.org)

#### *Background*

At SIF, Mr. Campbell is responsible for leading and taking ownership over providing value-add for members and potential members of the Association and their planned or ongoing SIF Implementation and Development. Prior to joining SIF, he was the Integration Team Manager for the Oklahoma State Department of Education. In his role, he managed many state and local projects aimed at improving the technology and data flow across 540 school districts and charter schools.

#### Key Take-away



- Virginia's SLDS model is definitely different from what other states have developed.
- Most current SIF development efforts are focusing on internal user portals.
- Federated model requires strong data governance.

#### *Summary*

Mr. Campbell has exposure to a number of LDS implementations around the nation. In all known implementations, he stated that Virginia's SLDS model is definitely different from what other states have developed. He further stated that SIF would be interested in the Virginia solution as SIF is concerned about integration across states.

The SLDS portal was described to Mr. Campbell as a public facing and internal facing implementation. He responded that most SIF development efforts are focusing on the internal user portal; however, the intent is to have one portal for both internal and external users. The requirements on data access are easier to manage for the internal users initially. When adding a public facing connection, the requirements become more difficult.

Towards the end of the interview, the team discussed data governance with Mr. Campbell. He explained that the federated architecture requires strong data governance as the model promotes a hierarchy of governance. He advised the team that the State of Washington may be a good model to investigate for data governance ideas.

## 5.4 Susan Carter

Title: Managing Partner  
 Organization: Data Strategies, Inc.  
 Phone: 804-965-0003  
 Email: [SCarter@DataStrategiesInc.com](mailto:SCarter@DataStrategiesInc.com)

### *Background*

Susan Carter has over twenty years of experience in the Data Management and Information Technology field. Ms. Carter built a successful Women-Owned Small Business (WOSB) in Data Management and IT consulting. Responsible for the research and integration of new technologies into several large corporations, she has helped organizations such as the Defense Logistics Agency (DLA), MCI (WorldCom), E.I. DuPont, and SmithKline Beecham gain increased efficiencies and market share through innovative uses of new and proven technologies.

### Key Take-away



- Multiple personal data elements can be used in combination to create a hash key that will result in more unique IDs.
- Use of multiple hash keys based on various sets of identifiable information allows different databases with different identifiable information to have a higher probability of forming linkages.
- The inability to store linkages will greatly limit performance
- Any changes to the data sources will require tight data governance

### *Summary*

Ms. Carter found Virginia's unique privacy requirements for the SLDS similar to data integration work she performed for the Army. This work combined information from various databases including military, financial, medical, and psychological. Similarly, the subject of interest for forming these data linkages was one that would require the utmost security to ensure the privacy of the individuals.

To address this issue, Ms. Carter suggested the use of not a single data element to create a hash key, but of multiple data elements that could be combined in a known way prior to being hashed. Another technique she suggested was the use of multiple hash keys based on a variety of separate or combined data elements. This technique would increase the possibility of creating matches across databases that may not have the same data elements across all databases, but have at least one. The other advantage to utilizing multiple hash keys as identifiers is the ability to match records when there are errors or inconsistencies in certain data elements such as a misspelled name in one particular database. The use of multiple hash key identifiers and an algorithm to determine confidence in a positive match would result in matches that a single hash key identifier would overlook. The key to successful matches would be through the identification of which data elements, or combinations, would result in the highest number of positive matches while reducing false matches. Ms. Carter suggested that the intelligence agencies (e.g.,

Department of Homeland Security, CIA, and FBI) are working intensively in the area of matching multiple identifiable elements and the development of algorithms that result in high confidence levels of matching.

Ms. Carter also spoke briefly on Virginia's inability to store linkages persistently. In her experience this would greatly limit the performance of a system. Ms. Carter has worked with some systems that have been designed to not store linkages due to privacy or other policy requirements; the system was able to adopt this design because there was a high level of importance on privacy and a low level of importance on performance requirements.

Lastly, the topic of data governance was touched upon during the interview. Ms. Carter stressed the importance of tight data governance in any system whose data sources would regularly have changes or when new data sources are added.

## 5.5 Raj Ramesh

Title: CEO  
 Organization: CTEC  
 Phone: 703-766-5774  
 Email: [rraj@ctec-corp.com](mailto:rraj@ctec-corp.com)

### *Background*

Mr. Ramesh has over 17 years of management and technical experience in information technology and software product development. Mr. Ramesh's areas of technical expertise include: Enterprise Database Systems/eLearning Systems, Data Warehousing and Very Large Database Systems (VLDB), Web-based Collaborative E-Learning Systems, Enterprise Architectures, Portals and CRM. Mr. Ramesh is a co-inventor of a Patent-Pending Task Share platform for the desktop environment and hand-held wireless devices.

### Key Take-away



- Used a three-tier design process for projects; presentation, business logic, database.
- In data analysis projects with large datasets, users accept delayed results.
- Development of global schema helps identify data type discrepancies.
- Data integration on-the-fly is a new technology area that has few products.

### *Summary*

Mr. Ramesh, like Mr. Carney, has extensive background with large scale system integration. His current projects involve multi-source data integration across disparate organizations (Federal Government). He uses a three-tier design process for projects; presentation, business logic, database. Mr. Ramesh's current projects consist of large data sets (tens of millions of records). In these data analysis projects, users accept long delays in receiving results. As the projects have moved to production and have matured, users now are requesting real-time capability in reporting.

When working with large disparate data sets, Mr. Ramesh stated that the development of a global schema helps identify data type discrepancies. There are visual mapping tools that provide good assistance in working out the global schema. The final point he made was that to his knowledge data integration on-the-fly is a new technology area that has few products.

## 5.6 Ron Kleinman

Title: CTO  
 Organization: SIF Association  
 Phone: 202-607-8526  
 Email: [rkleinman@sifassociation.org](mailto:rkleinman@sifassociation.org)

### *Background*

Prior to joining SIF, Mr. Kleinman was the Chief Technical Evangelist for Sun Developer Relations, and served as Sun's representative on multiple industry-wide Java and XML standards committees. He has extensive experience consulting with developers who are trying to "java-tize" their existing applications. He has prepared and delivered numerous presentations on Java technologies both in the U.S and overseas. His particular areas of expertise include Java on the Server (EJBs and server-side APIs), Jini, Java-based device access control and management, and more recently, XML.

### Key Take-away



- Granularity of searching can impact security policy implementation.
- Real-time hashing could be a performance anchor.
- Use of federated identity concepts may provide a solution for record mapping across sources.
- Use of a central security authority built into the process could enforce strong data protection in the system.
- Separation of cross-walk table with firewall.

### *Summary*

Dr. Kleinman spent a majority of the interview focused on the security implementation for the LDS. He pointed out that the granularity of searching can impact security policy implementation. The more the implementation allows finer resolution in the search, the stronger the security policy needs to be defined. Dr. Kleinman noted that the use of a central security authority could enforce strong data protection in the system.

Dr. Kleinman recommended that the team consider federated identity concepts to provide a solution for record mapping across sources. This federated identity can be used to enforce consistent security policies across the data sources. He felt that security measures provided by operating systems, database management systems, firewalls, and routers can add to the security implementation for the LDS. For example, the linking/crosswalk table in the design could be separated from the rest of the system with a firewall.

A final point made by Dr. Kleinman was the concern that real-time hashing could be a performance anchor for the LDS. Some users may find the delay in results to be unacceptable.

## 5.7 Peter Dobler

Title: President  
Organization: Dobler Consulting  
Phone: 813-322-3240  
Email: [pdobler@doblerconsulting.com](mailto:pdobler@doblerconsulting.com)

### *Background*

Mr. Dobler started his professional career more than twenty two years ago in software development. After working many years as a consultant for the three largest Swiss banks he founded his own consulting business in 1997. Mr. Dobler is a recognized expert in Sybase ASE, Sybase Replication Server and Sybase IQ. He also has many years of Oracle experience, including the latest 11g release. Mr. Dobler also has in-depth knowledge of SQL Server 2000 and 2005.

### Key Take-away



- The problem with a federated database is the performance
- Create a hash using not only identifiable information but external data as well, such as the source.

### *Summary*

The majority of the discussion with Mr. Dobler focused on the performance of the federated database and how to improve its efficiency. If Virginia were able to perform various filtering of the data prior to entering the SLDS's data engine, the performance could be improved significantly. He suggested off-the-shelf solutions including those from Sybase IQ and SAP.

Mr. Dobler also weighed in on the use of hash key identifiers and suggested a hash key based on a combination of data element(s) and external data. In a particular example, he combined PII data within a record with the source database identifier. The use of external data would help to add a layer of protection and information to the hash key.

## 5.8 Dr. Laura Haas

Title: IBM Fellow  
 Director, Computer Science  
 Organization: IBM Almaden Research Center  
 Phone: 408-927-1700  
 Email: [laura@almaden.ibm.com](mailto:laura@almaden.ibm.com)

### Background

Dr. Haas is an IBM Distinguished Engineer and Director of Computer Science at Almaden Research Center. Previously, Dr. Haas was a research staff member and manager at Almaden. She is best known for her work on the Starburst query processor (from which DB2 UDB was developed), on Garlic, a system which allowed federation of heterogeneous data sources, and on Clio, the first semi-automatic tool for heterogeneous schema mapping. Dr. Haas is Vice President of the VLDB Board of Trustees, a member of the IBM Academy of Technology, and an ACM Fellow.

### Key Take-away



- When performing joins on-the-fly, it is important to minimize the volume of the data and minimize the trips back and forth.
- A semi-join can be an efficient way to link data on-the-fly and may be better than a linking table/directory.
- A nested-loop join can be used on the most constrained source from the query and then join data from the other sources as needed.
- Commercial databases tend not to use join indexes due to the amount of look-ups which degrade performance.

### Summary

The interview with Dr. Haas was one focused on optimizing the desired architecture of Virginia's SLDS. Knowing that Virginia would not be able to store linkages persistently and the system would need to be a federated database, Dr. Haas was able to share various techniques that would improve the performance of the system. First and foremost, Virginia would need to minimize both the volume of data being transmitted and minimize the back and forth trips between the data sources and the SLDS.

Instead of a linking table, Dr. Haas stated that in-memory joins would likely result in greater performance, depending on the types of queries that are conducted. Linking tables or join indexes are not typical of commercial databases as they could require multiple look-ups, thereby wasting cycles and decreasing efficiency. Dr. Haas did not suggest a linking table for multiple database systems. The SLDS' architecture could be designed to maximize performance by identifying the types of queries and the frequency pattern of each type of query. Semi joins, merge joins, and nested-loop joins were different types of joins that Dr. Haas recommended as viable options, each with performance benefits depending on the database size, query type, and query frequency patterns.

## 5.9 *Dr. Thilini Ariyachandra*

Title: Assistant Professor, Management Information Systems  
 Organization: Xavier University  
 Phone: 513-745-3379  
 Email: [ariyachandrat@xavier.edu](mailto:ariyachandrat@xavier.edu)

### *Background*

At Xavier, Dr. Ariyachandra teaches principles of information systems, business intelligence and data management and modeling exposing students to BI and database offerings by Teradata, Oracle, Microsoft and Microstrategy. She has received several awards for scholarly excellence. Her research is focused on the selection, design, implementation of business intelligence solutions in organizations, information system success as well as impacts of social networking. She has published in high impact practitioner and academic journals.

### Key Take-away



- In a federated model, small query sets work best, and ad hoc can lead to poor performance.
- Trends show distributed database implementations moving towards the federated model.
- In-memory databases are not as scalable as standard DBMS.

### *Summary*

Dr. Ariyachandra's paper that compared different data warehouse architectures provided a starting foundation for the interview. She actually was hoping to talk the team out of the federated model, but soon realized that was not an option. In a federated model, Dr. Ariyachandra noted that a small query sets work best (controlled environment), and ad hoc can lead to poor performance. Based on her interpretation of the architecture, the proposed federated architecture does not easily support ad hoc analysis of data and, therefore, may not be impacted by performance.

Dr. Ariyachandra explained that her current consulting engagements show that trends in distributed database implementations are moving towards the federated model. She also noted that in-memory databases are not as scalable as standard DBMS; however, vendors are starting to offer product options.

## 5.10 Dr. Cynthia Dwork

Title: Distinguished Scientist  
 Organization: Microsoft Research  
 Phone: 650-693-3701  
 Email: [dwork@microsoft.com](mailto:dwork@microsoft.com)

### *Background*

Dr. Dwork is the world's foremost expert on placing privacy-preserving data analysis on a mathematically rigorous foundation. A cornerstone of this work is differential privacy, a strong privacy guarantee permitting highly accurate data analysis. Dr. Dwork has also made seminal contributions in cryptography and distributed computing, and is a recipient of the Edsger W. Dijkstra Prize, recognizing some of her earliest work establishing the pillars on which every fault-tolerant system has been built for decades.

### Key Take-away



- There's no principal way to sanitize data.
- Re-identification techniques are getting faster and cheaper.
- Differencing attacks are able to re-identify information even with large aggregate data.
- The addition of “noise” is used to prevent various attacks including differencing and averaging.
- VDOE should consider whether historical records should be archived after a period of time as a security measure.

### *Summary*

As a privacy expert, Dr. Dwork's interest in Virginia's SLDS was the security rules being placed on the architecture and on the roles of users given access to the database. She was quick to state that re-identification of data was becoming faster and cheaper and, as a result, that there was no principal way to sanitize data. Dr. Dwork stressed the importance of data security and anonymization. She offered her insights on various techniques that are used to re-identify data and countermeasures to those techniques.

Upon learning that Virginia intended to present aggregate data to the public, Dr. Dwork provided examples of “differencing attacks” that use two large data sets with similar information and could be used to re-identify individuals. To combat these differencing attacks the introduction of data “noise” can be added to the actual data which, if done properly, can effectively stop the ability to average the results of aggregate data. The addition of noise should not significantly skew the aggregate data, but may be against certain regulations or policies. Another means of protecting against differencing attacks is to limit a user's ability to submit queries or limit their ability to run queries that are too similar.

As a general question, Dr. Dwork asked if the SLDS would have a limit on its ability to query historical data. By archiving or limiting the SLDS' access to data older than a defined period, the SLDS can again limit potential security attacks. This limiting of data also would have the added

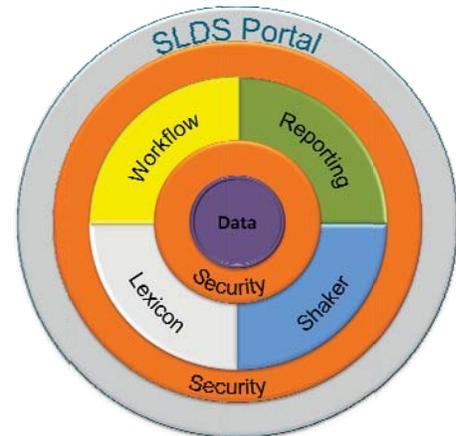
benefit of reducing the total records within the database which would increase performance of the system.

When discussing the internal use of the SLDS and the ability of researchers to gain access to record level data, Dr. Dwork talked about the high risk of privacy breach and the need for limitations and rules that should be placed on these users. Security of a system must be implemented as a combination of both technological security measures and security policy that governs the individuals with access to the system; only by having both, can a system maintain a high level of security of its information.

## 6 SLDS Architecture Overview

The SLDS architecture consists of seven functional components. Commercial-of-the-shelf (COTS) products will be used where applicable, and shared computing resources will be used in the physical implementation where applicable.

The SLDS architecture can be represented by a bull's-eye signifying the data-centric nature of the architecture. The importance of security is reflected in the representation through the dual rings that surround critical components located inside the portal. For example, a security ring surrounding data indicates tight security of that data component, and the ring surrounding the four tools and task oriented components illustrates security controls built into the other functional components. The SLDS Portal provides the key interface into the architecture.



### 6.1 SLDS Seven Functional Components

#### 6.1.1 Portal

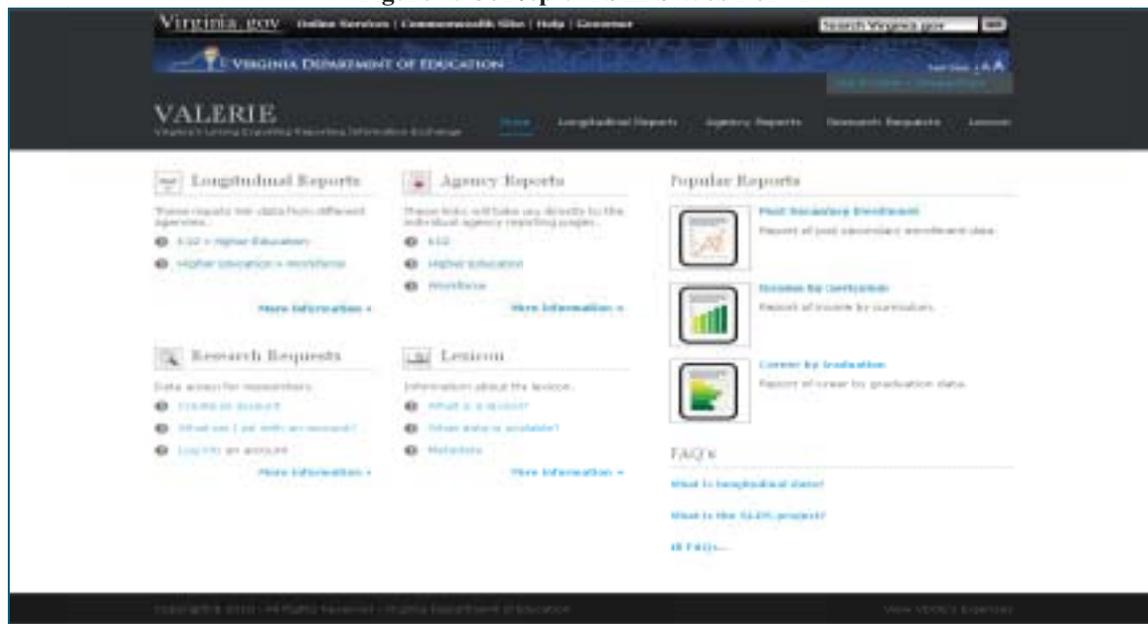


The front door into the Statewide Longitudinal Data System (SLDS) is through the SLDS Portal. The SLDS Portal provides both public (*anonymous*) and private (*named*) users with a variety of functions and services. Development of the Portal will be performed using a modern application framework, e.g., .Net or Java and a content management system, e.g., DotNetNuke or Umbraco.

Named users gain access after they have requested an account, and their request has been approved by the appropriate agencies. Once approved, the named user account has access to help, training, the Lexicon, requests for data, status of requests, and account maintenance including password reset.

#### *SLDS Portal Components*

The Portal provides access to virtually all of the SLDS components to include the Shaker, Reports, Lexicon, Data, and a limited amount of Workflow. In addition to the SLDS Components, the Portal provides services such as help files, frequently asked questions (FAQs), hyperlinks to Agency reports, and the ability to request a private (*named*) account. Figure 4 provides a conceptual representation of the functions and services which will be accessible through the SLDS Portal.

**Figure 4: Conceptual SLDS Web Portal**

### *Public (Anonymous) User Functionality*

Public user will have access to the following features and functionality:

- Help files for functions which are available to public users.
- Frequently asked questions (FAQs).
- Prebuilt aggregated data reports which have been approved by the governance structure.
- Lexicon elements which have been approved by the governance structure.
- Hyperlinks to Agency reports on other websites.
- Electronic request and workflow for Named User Account requests.

### *Private (Named) User Functionality*

Private users will have access to additional functionality which is not available to public users. Requests for private accounts will be submitted electronically using elements of the SLDS Portal and the Workflow component. Procedures for submitting, approving and denying account requests will be delineated by the governance structure.

Users who have been approved for a private account will be notified by email. Access to private account features will be granted after users have supplied a valid username and password and have been authenticated against either the COV or COV AUTH directory. Private users may, depending on their permissions, have access to the following features and functionality:

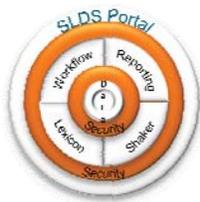
- Help/Training files to include “How To” and instructional videos.
- Reports
  - Ability to view non-suppressed aggregated.
  - Ability to access the Query Building Tool (QBT) for constructing data requests.
- Lexicon
  - Functionality determined by the governance structure.
- Workflow

- Ability to electronically submit and track Data request.
- Ability to retrieve data which has been requested and approved.
- Ability to attach files.
- Ability to check status, modify or cancel account and/or data request.
- Password reset
  - Ability to reset the user’s password. This capability may be provided through the COV AUTH directory process.

### *Application Framework and Content Management System Features*

The SLDS Portal will be developed using a modern application development framework and content management system (CMS). Use of an application framework, such as Microsoft .Net, and a content management system, such as Umbraco, allows for the development of rich functionality and services with minimal development. Most content management systems include features and services such as web URL control, custom content types and views, revision control, taxonomy, user management, documentation, and established community support.

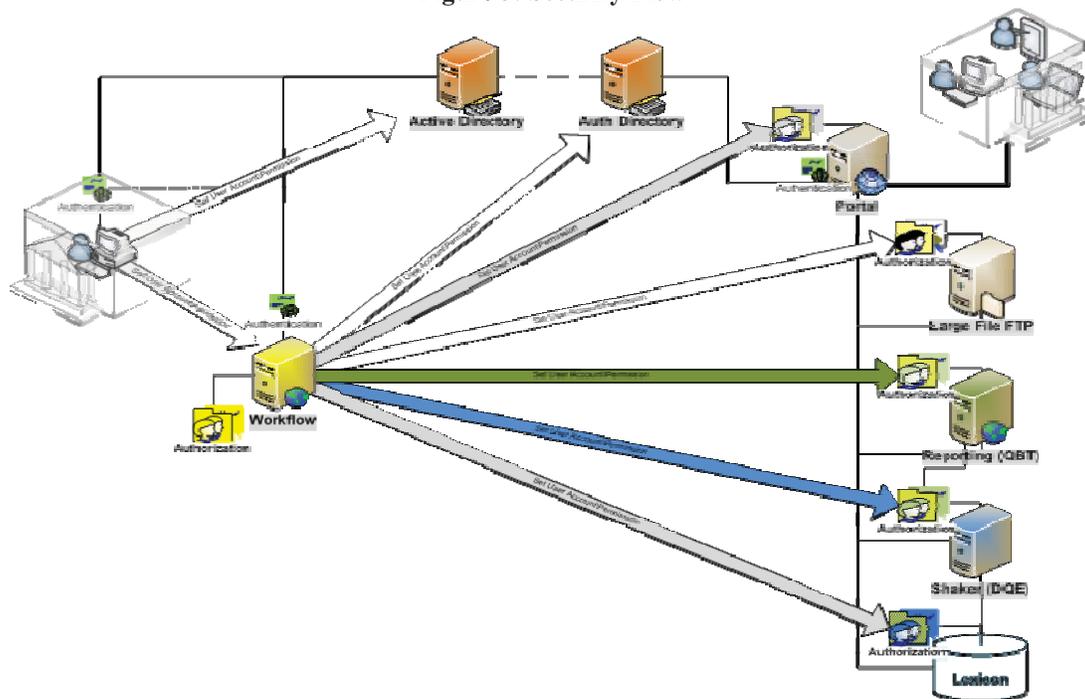
#### 6.1.2 Security



Security is the foundation component for the SLDS. The sensitivity of the information and policies regarding who and how data is handled will be managed through a cohesive security model. The model used for the SLDS incorporates authentication and authorization pieces.

Authentication is required for all private (*named*) users, to include researchers as well as agency employees. Researchers and agency employees will be authenticated as a precondition to gaining access to the named user portions of the SLDS portal. Agency employees will be authenticated before gaining access to the Workflow component of the SLDS application. Figure 5 depicts the interaction of the Workflow component with other SLDS components, as well as the authentication interface for agency employees and researchers.

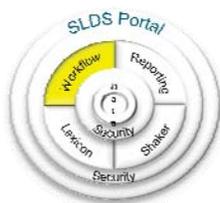
When requests for accounts and data access are submitted through the SLDS Portal, the Workflow component triggers messages to designated Commonwealth of Virginia (COV) employees for review and action. The action takes the form of approval or denial. In order for a COV employee to interact with the Workflow component, s/he would need to log in (authenticate by the COV Active Directory) to the COV infrastructure. Thereafter, s/he would be able to access the Workflow component in order to act on the Workflow trigger.

**Figure 5: Security Flow**

For researchers, authentication would occur through the SLDS Portal using the COV AUTH directory. After an account request is approved, a researcher would be required to log in to their accounts to make data requests from resources for which they have received approval. Authorization defines user roles and the permissions associated with those roles.

For example, a researcher (role) would have access to view (permission) the Lexicon, while a data administrator (role) would have access to view and modify (permission) the Lexicon. The Workflow component is the hub for managing a user's roles and associated permissions. The SLDS components coordinate with the Workflow to manage requests for services correctly.

### 6.1.3 Workflow

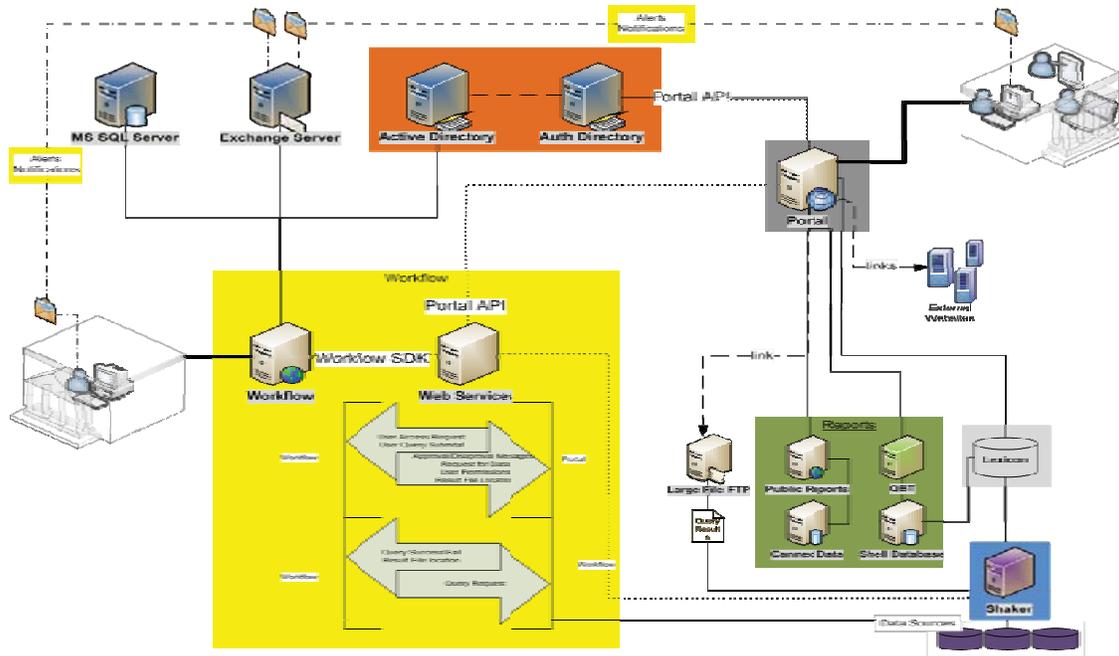


The back office of the SLDS is the Workflow component. The SLDS Workflow will be developed using the Microsoft Dynamics Customer Relationship Management (CRM) package. CRM is a solution for automating internal business processes by creating workflow rules that describe routine tasks involving daily business operations. These processes can be designed to make sure that appropriate and timely information is sent to the correct people. To initiate workflows and in order to act upon the successful completion of workflows the SLDS Workflow component will need to have interfaces to the other components of the SLDS Portal.

The main function of the SLDS Workflow component is to manage and define a series of tasks within an organization to produce a final outcome or outcomes. These workflows will allow the partner agencies to work together to control access to their shared data and system. The workflows will handle email alerts and notifications to both the partner agencies and to the

portal named users. Figure 6 provides a conceptual representation of the architecture and interfaces of the SLDS Workflow Component.

**Figure 6: Workflow Component**



### Interfaces

To initiate workflows and in order to act upon the successful completion of workflows the SLDS Workflow component will need to have interfaces to the other components of the SLDS Portal.

#### Portal Interfaces

Named users will interact with the SLDS Portal in order to submit data to the Workflow component to initiate the following workflow processes:

- User Access Request
- User Query Request

Once the workflows have been completed, the result of the process will be communicated back to the named user through the Portal. The Workflow component will need to be able to push back the following information to the SLDS Portal:

- Approval/Disapproval messages,
- Request for additional data,
- Portal user role permissions,
- And query result file location.

#### Shaker Interfaces

When a query request is approved, the Workflow component will interact with the Shaker in order to submit the query for execution. The Shaker will notify the Workflow component of the success or failure of the query execution and, in the event of success, the resultant file location.

### *Multiple Component Interfaces*

The Workflow component will be used to coordinate the access authorization in multiple components of the SLDS system. This will allow user access to be centrally administered but distributed to the individual components based on purpose and need.

### *Workflows*

The SLDS Workflow component manages and defines a series of tasks within an organization to produce a final outcome or outcomes. It will allow the SLDS team to define different workflows for different types of jobs or processes. At each stage in the workflow, one individual or group is responsible for a specific task. Once the task is complete, the workflow software ensures that the individuals responsible for the next task are notified and receive the data they need to execute their stage of the process. It will also automate redundant tasks and ensure incomplete tasks are followed up.

### *User Account Request*

A workflow must exist to review a request for named researcher access to the Portal. The workflow will result in one of the following outcomes:

- Approve,
- Disapprove,
- Or request more data (update account request).

A workflow must exist to review a request for named data owner access to the Portal. The workflow will result in one of the following outcomes:

- Approve,
- Disapprove,
- Or request more data (update account request).

### *User Ad Hoc Query Request*

A workflow must exist to review a query request for a named researcher. The workflow will result in one of the following outcomes:

- Approve,
- Disapprove,
- Or request more data (update query request).

### *User Ad Hoc Query Result*

A function must exist to handle the result of a query request for a named researcher. The function must perform the following tasks:

- Receive back status,
- Receive file location,
- Communicate resultant file location to the named researcher through the portal,
- Communicate status to the named researcher through the portal,
- Communicate status to the named researcher through an alert or email,
- And communicate failed status to proper administrator.

All account and data requests are processed through and managed by the Workflow component. Workflow monitors and triggers actions such as query submission and maintains status of requests. Workflow is the source of information about roles and permissions for SLDS users.

When an account request is submitted, it is the Workflow component that manages the message(s) and notifies designated COV employees about the request. Through the Workflow component, employees can approve or deny the request. Workflow then notifies the submitter of the account request of the final decision.

On a data request, Workflow monitors the request, confirms approval, and submits the query to the Shaker for action. Designated COV employees are notified of the request to approve or deny the query. If the request is denied, Workflow notifies the researcher of the denial status. If approved, Workflow submits the request to the Shaker and continues to monitor status. Upon completion of the transaction, Workflow notifies the researcher the data set is available for download.

### 6.1.4 Reporting



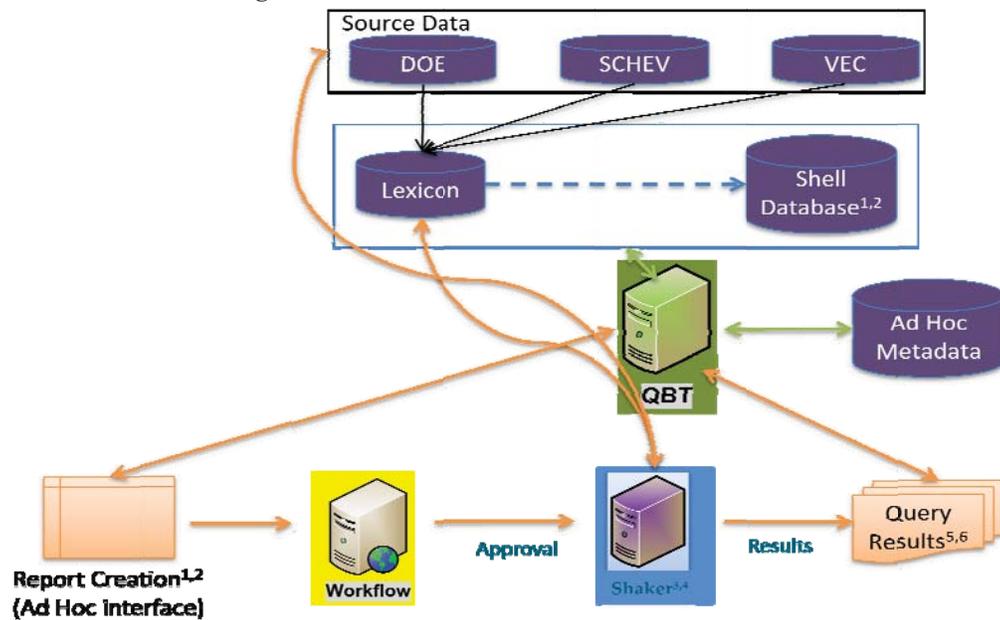
The SLDS Business Intelligence (BI) architecture will support two scenarios:

- Ad Hoc reports of record-level user data.
- Pre-defined (canned) reports of aggregated linked data.

#### *Ad Hoc Record-Level Data BI Architecture*

A visual representation of the Ad Hoc record-level BI architecture is presented in Figure 7. This architecture consists of the following major components:

- A Lexicon and Shell Database that are built based on the source data and will support the Logi Ad Hoc tool.
- Report Creation using the Logi Ad Hoc Business Intelligence platform. This consists of the Logi Ad Hoc Query Building Tool and Ad Hoc Metadata.
- A Workflow engine that routes report submissions through an approval process.
- The Shaker that will service the query against record level data in the disparate source systems.
- The record level Query Results that will be presented to the user.

**Figure 7: Ad Hoc Record-Level BI Architecture**

### *Lexicon and Shell Database*

The Lexicon will contain information about the data objects that each of the source data systems have made available to the SLDS. This information will be utilized to create a Shell Database. The Shell Database serves two purposes: (1) it is an instantiation of the Lexicon and is needed by Logi Ad Hoc in order to function; (2) it contains sample data that will be used by Logi Ad Hoc to allow researchers to preview their query request. A process will be built that creates and populates the Shell Database based on the information available in the Lexicon.

### *Report Creation*

The Report Creation process will be provided through Logi Ad Hoc. Logi Ad Hoc has a self-serve user interface to allow a user to specify a report query. It has facilities for the user to save the query and preview a sample of the report. When the user is finished with the report, they can run the report. During this step, a custom process will intercept the query that has been submitted. The query produced along with the parameters including columns selected and filters specified will be sent to the Workflow component.

### *Workflow*

The Workflow component routes the query through the appropriate steps to get acceptance by specified agency reviewers.

### *Shaker*

The Shaker interacts with the Lexicon and the source data systems to query and join together the record level data from each of the source data systems. It deposits the joined data set in either a file or a database table.

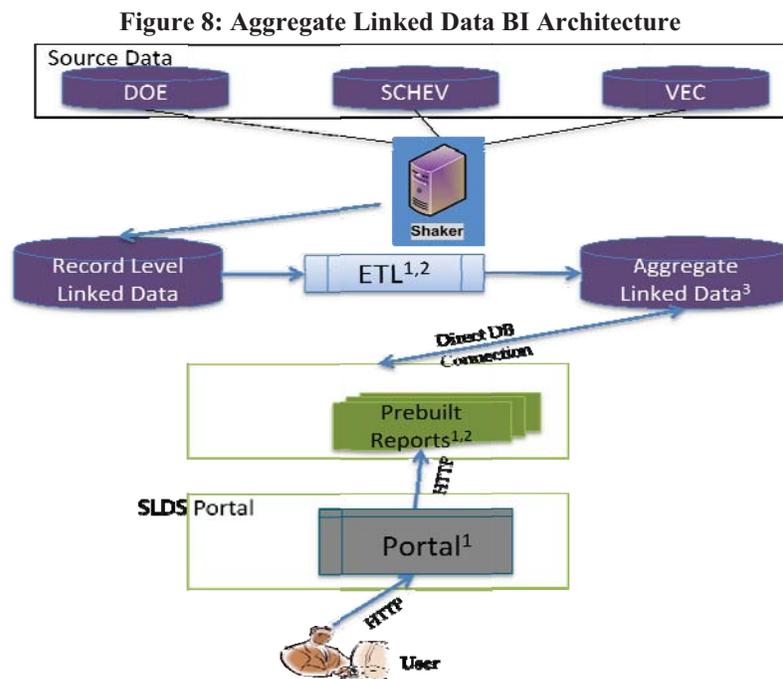
### Query Results

As mentioned previously, the Shaker has the option to place the resulting query data in a database table or file. If the results are sent to a file, the submitting researcher is notified that the file is available for download by the researcher from the Portal. If the results are placed in a database table, reports can dynamically be created in Logi Info that references the data in these tables. These reports will provide the user with some limited capabilities for analysis like filtering, sorting, and grouping of the data.

### Aggregated Linked Data BI Architecture

A visual representation of the BI architecture for aggregated linked data is presented in Figure 8. This architecture consists of the following major components:

- The Shaker to join record level data from the source data systems.
- A repository for the record level linked data.
- An ETL (extract, transform, load) tool and process to extract data from the record level linked data store and load it into the aggregated linked data store.
- A repository for the aggregated linked data.
- The Logi Info Business Intelligence platform that will be used for serving up prebuilt reports.
- The SLDS Portal where the Logi Info reports will be embedded.



### Shaker

The Shaker interacts with the Lexicon and the source data systems to query and join together the record level data. It deposits the joined data set into a database table in the Record Level Linked Data Store.

### *Record Level Linked Data Store/ETL Process/Aggregate Linked Data Store*

An ETL process takes data from the Record Level Linked Data Store, aggregates it to the appropriate level, and loads the aggregated data into the Aggregate Linked Data Store. Once the data has been loaded, the table in the Record Level Linked Data Store will be purged.

### *Prebuilt Reports*

Prebuilt Reports will be created using Logi Info and will use data in the Aggregate Linked Data Store. The Logi Info product allows for custom design and format reports that pull data from a pre-specified data source. The reports produced can contain tables, charts, and maps or a combination. Logi Info has some prebuilt analytical type reports that allow an end user to perform some limited analysis of the data including sorting, filtering, and grouping the data.

### *SLDS Portal*

The Prebuilt Reports will be made available in the SLDS Portal.

## 6.1.5 Lexicon



*“The Lexicon is an inventory of every available data field in every available data source, the structure of their storage, the possible values and meanings of the information stored, all possible transformations of each set of field values to another set of field values, methods of data source access, and matching algorithms and how they are to be used in conjunction with possible field value transformations.”*

The Lexicon (Figure 9) contains no data from any data source. It will be used to manage the Shell Database for users to build queries against, as well as providing the Shaker with appropriate information to prepare an optimized query sequence for data requests. The Shell Database will contain fictitious data.

A researcher, when building a query, interacts with a set of field names and relationships to formulate a query. The user interface for the query building provides a simple view of the Lexicon for easy query construction.

To maintain the accuracy and to manage extensibility of the Lexicon, the component processes all data sources periodically at a predetermined time/interval searching for:

- Changes in data ranges,
- new data fields,
- and anything else that would disrupt the probabilistic matching or provide more ways to “slice and dice” the data.

Anomalies found by the linking module will prompt an alert for an administrator to modify the matching algorithm or add new query choices.

**Figure 9: A Logical Representation of the Lexicon and its Interactions**

### 6.1.6 Shaker



The Shaker's general function (Figure 10) is to accept an approved query and return a dataset. The query will be broken down into a series of optimized steps, or sub-queries, to retrieve de-identified data from the appropriate data sources in the most efficient manner. In keeping with the intent of the original request, query forms (e.g. inner join, left join, equijoin) and specified final output parameters (e.g. counts of non-matching records by demographic categories) will be taken into consideration. Information from the Lexicon concerning data structure and relationships will be used to produce a dynamic sub-query plan for data retrieval that minimizes processing time and workload on the target data sources.

For each query submitted to the Shaker, a random key is generated. Each sub-query in the data retrieval plan will send this random key to the data source to be used in creating a secure one-way hashed key for any applicable records. This list of hashed keys is then used by the Shaker to combine records across multiple data sources, never transferring any identifiable information out of the data source. Any hashed keys used to link records will be removed from the final data set and replaced with yet another random key which cannot be traced back to any original data sources. The resulting combined records are then uploaded to a file or database table for later access by the user.

**Figure 10: A Logical Representation of the Shaker and its Interactions**

### 6.1.7 Data



This SLDS Data Architecture (Figure 11) consists of the Source Data Systems. The Shaker will submit queries to the target data systems and join the resulting data sets. The data will optionally be written to a database that resides in the SLDS environment. The SLDS environment will also contain other databases needed by the SLDS Portal including a Metadata/Security database, Workflow database, Lexicon, Shell database, and an Aggregate Linked Data database.

#### *SLDS Databases*

Several databases will reside in SLDS environment. These databases will act as a repository for data and metadata needed by various components of the SLDS Portal. Each database and its Portal usage are described below.

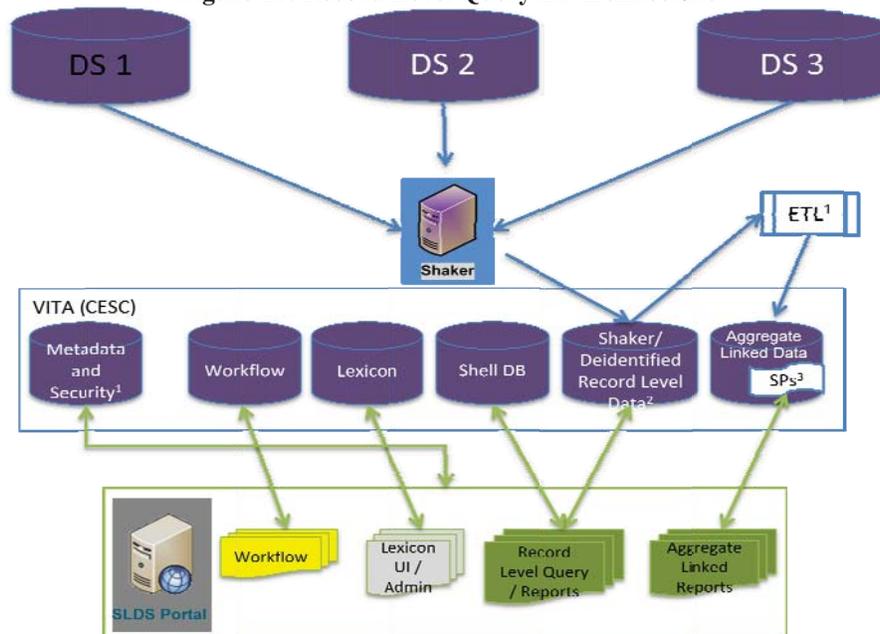
#### *Metadata and Security Database*

The Metadata and Security database will consist of the Logi Ad Hoc metadata, logging data repository, auditing data repository and any data that needs to be maintained to control security for the portal.

#### *Workflow Database*

The Workflow database will contain data needed by the Workflow engine. This will include data needed to track the steps and processes in the workflow. It will also include data required for security to be maintained in the workflow.

**Figure 11: Record Level Query BI Architecture**



### *Lexicon*

The Lexicon database will provide information about the data objects that have been exposed to the SLDS Portal by each of the source data systems. Interaction with the data store will be through the Lexicon user interface and administration portion of the SLDS Portal.

### *Shell Database*

A Shell Database is needed in order for Query Building Tool to function. The Shell Database will be built off of information contained in the Lexicon.

### *Shaker/De-identified Record Level Linked Database*

The Shaker will optionally write the results of the joined data from the Source Data Systems to the De-identified Record Level Linked Database.

### *Aggregate Linked Data*

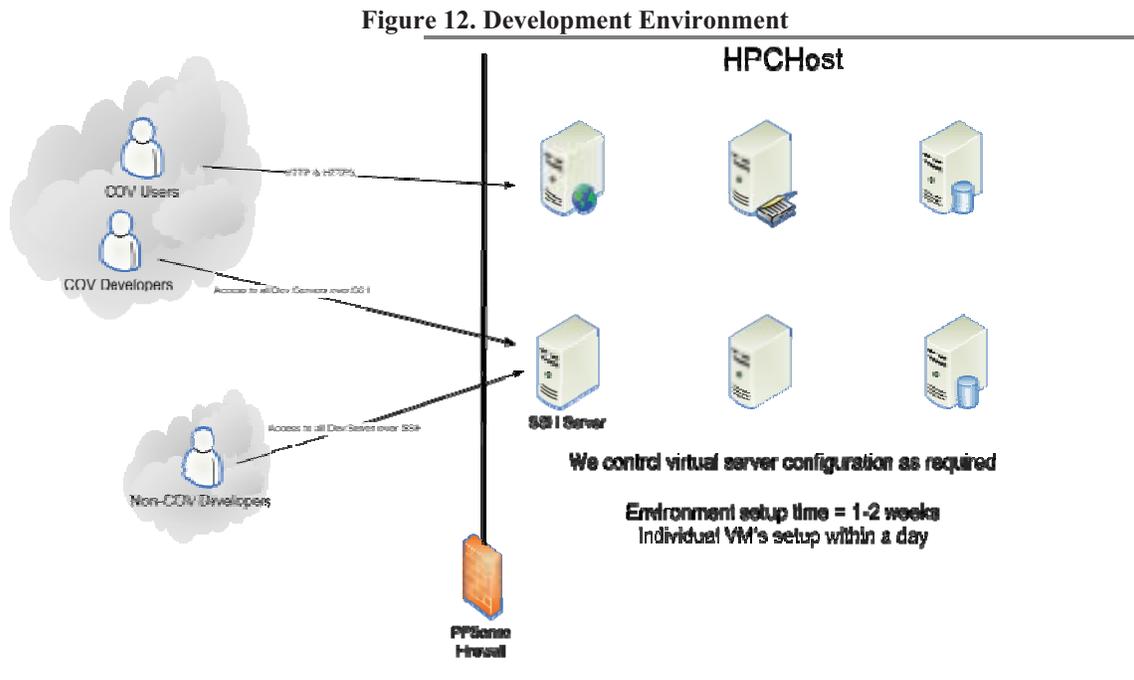
The Aggregate Linked Database will be utilized by the prebuilt reports. This database will be populated through ETL processes that will aggregate data from the De-identified Record Level Linked Database. Stored procedures will be used by these reports for data querying and suppression.

## 7 Physical Infrastructure

The SLDS application will be developed, tested and deployed in three environments. The Development environment will be hosted at a Florida datacenter operated by HPCHost.com. The Test and Production environments will be hosted at the Commonwealth Enterprise Solutions Center (CESC) physically located in Chesterfield Virginia.

### 7.1 Development Environment

The Development environment (Figure 12) will be purchased as a monthly service. The environment will consist of Virtual Machines (VM) operating on HPCHost managed VMWare ESX infrastructure. This monthly service can be increased, decreased or eliminated as necessary.



A request has been submitted to purchase a “slice” of computing power which consists of a 1 4-Core processor, 16 GB RAM and 300 GB of RAID 6 SAN storage. VITA EAD will manage the Development VM’s using the following rules/guidelines:

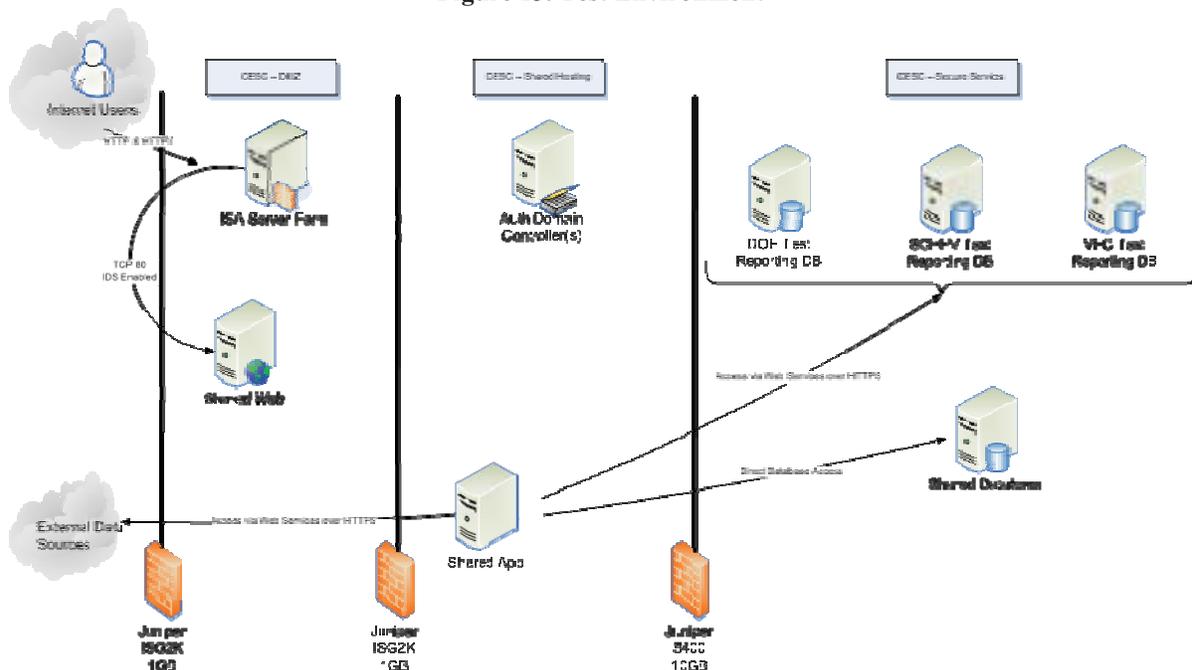
- VITA EAD will build/configure each VM as requested by the developers.
- It is estimated that 6 – 12 VM’s will be built for this development effort.
  - Portal Server
  - Report Server
  - Workflow Server
  - Shaker Server
  - Database Servers (2)
- All Microsoft OS based VM’s will be joined to the VITA EAD operated EADDEV Domain unless otherwise requested.

- VITA EAD will operate a PfSense Firewall that will protect the Development environment.
- Developers will access the environment using a VPN or a SSH tunnel.
- Non-Developers on the COV network may be allowed web access to the development portal and reports as required.
- Developers will be given full administrative control over their respective VM's.
- All Code will be checked into a Code Repository operated by VITA EAD.
- Development VM's will not be backed up unless specifically requested.

## 7.2 Test Environment

There will be no additional servers specifically ordered for the SLDS Test environment (Figure 13). VITA EAD has ordered new physical servers that will have the capacity to support SLDS testing along with other applications. This methodology was adopted to reduce infrastructure costs by sharing physical infrastructure. VITA EAD will be responsible for all applications residing on these shared servers.

**Figure 13. Test Environment**



The test web and application servers will have two 12-Core processors, 128 GB RAM and 144 GB of RAID 5 local disk space while the test database server will have two 4-Core processors, 32 GB RAM and 273 GB of RAID 5 local disk space.

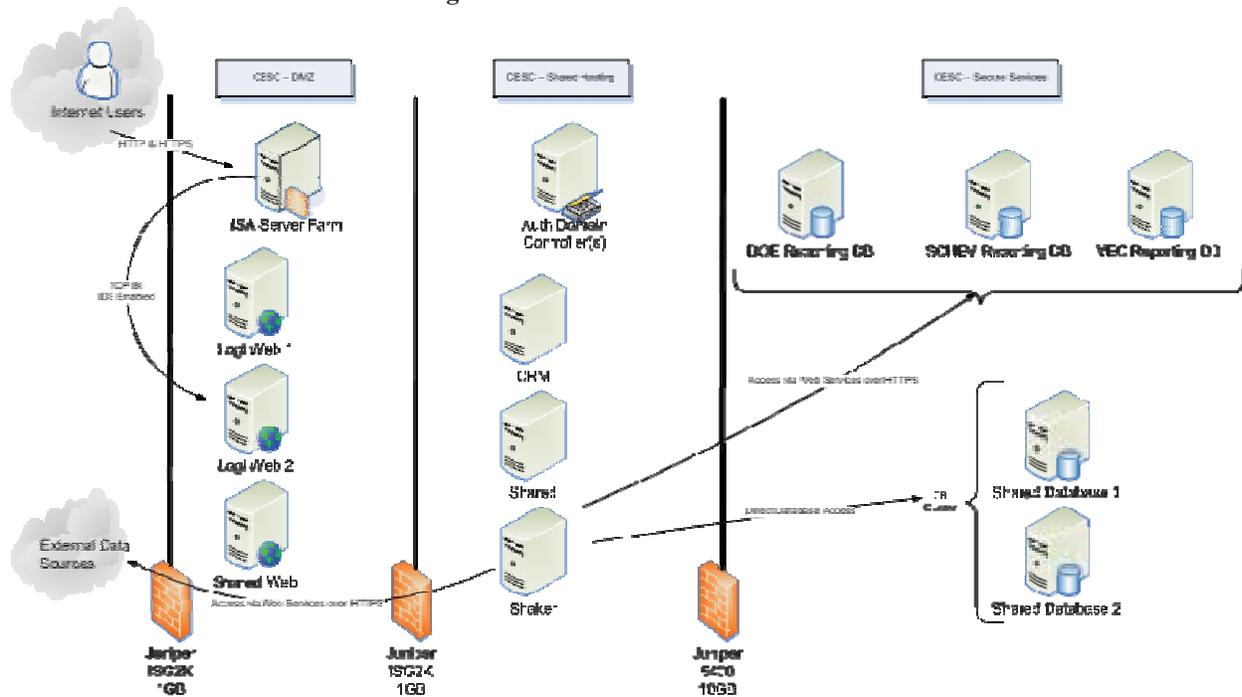
The SLDS application will leverage DOE and SCHEV test databases in their test environments. The SLDS application will primarily use web services to connect and will use that same method to connect to any future internal or external data source.

The Test environment web servers will be accessible from the Internet. VITA EAD system administrators will access the backend SLDS servers via 2-factor VPN RDP. DOE and SCHEV will control administrative access to/from their test databases. Developers will normally not have access to the SLDS test servers. Code/changes from the SLDS Development environment will be promoted to the SLDS Test environment through a structured promotion process.

### 7.3 Production Environment

An additional web and application server will be required for the SLDS Production environment (Figure 14). SLDS will also share new VITA EAD production servers. VITA EAD has ordered new physical servers that will have the capacity to support various SLDS components along with other applications. This methodology was adopted to reduce infrastructure costs by sharing physical infrastructure. VITA EAD will be responsible for all applications residing on the SLDS and shared servers.

Figure 14. Production Environment



All SLDS and shared servers will have two 12-Core processors, 128 GB RAM and 144 GB of RAID 5 local disk space. The SLDS application will leverage DOE and SCHEV production databases in their production environments. The SLDS application will primarily use web services to connect and will use that same method to connect to any future internal or external data source.

The Production environment web servers will be accessible from the Internet. VITA EAD system administrators will access the backend SLDS servers via 2-factor VPN RDP. DOE and SCHEV will control administrative access to/from their production databases. Developers will not have access to the SLDS production servers. All test and production servers will be joined to the COV domain. Code/changes from the SLDS Test environment will be promoted to the SLDS Production environment through a structured promotion process.

## Appendix A: Secondary Architecture Best Practice Case Studies

### A.1 Illinois State Board of Education

State/Agency: Illinois State Board of Education  
 Web Site: <http://www.isbe.state.il.us/ILDS/htmls/project.htm>  
 Address: 100 North First Street  
 Springfield, Illinois 36104  
 POC: Michael McKindles  
 POC Phone: 217-782-0329  
 POC Email: [mmckindl@isbe.net](mailto:mmckindl@isbe.net)

Case Profile
<b>Student Enrollment:</b> 2,119,707 <sup>35</sup>
<b>Teachers:</b> 135,704 <sup>36</sup>
<b>LDS Grant:</b> \$11,869,819 <sup>37</sup>

#### *Background*

In July 2009, Illinois Governor, Pat Quinn, signed into law the P-20 Longitudinal Education Data System Act<sup>38</sup>. The act was a response to the 2009 LDS grant Illinois received from the U.S. Department of Education. The grant funded the Illinois Longitudinal Data System (ILDS), which is being built to establish the technical and management systems necessary for the Illinois Board of Education (ISBE) and its education partners to manage, link and analyze P-20 education data.

In December 2009, the ISBE created the Illinois Data System Advisory Committee. At its inception, members included the Assistant Superintendent, the Division Administrator, and the ILDS Project Manager.

#### *System Design and Architecture*

Prior to its receipt of the federal LDS grant, ISBE implemented a state Student Identification System (SIS) and expanded its use. As of 2009, the ISBE SIS included five years of student enrollment data and program information; updated student demographic information; and four years of assessment results. The various data sources provide data on teacher demographic, teacher certification, LEA and school program participation, LEA financial information, LEA

<sup>35</sup> State educational data profiles. (n.d.). Retrieved from <http://nces.ed.gov/programs/stateprofiles/sresult.asp?mode=short&s1=17>

<sup>36</sup> Ibid.

<sup>37</sup> Statewide longitudinal data system grant program - grantee state - Illinois. (n.d.). Retrieved from <http://nces.ed.gov/programs/slds/state.asp?stateabbr=IL>

<sup>38</sup> Illinois Public Act 96-0107. (n.d.) Retrieved from <http://ilga.gov/legislation/publicacts/96/096-0107.htm>

facilities, specialized student programs, LEA compliance and monitoring, and LEA child nutrition services. As a result, ISBE had more than 100 disparate collection systems on a range of technologies.

The ISBE data system currently cannot provide data that can be used effectively in education decision making. Data currently collected by the agency is highly fragmented across various systems and collection vehicles. This fragmentation covers multiple data systems that include student level data, as well as a variety of systems that maintain data from other parts of the ISBE education enterprise (e.g. staff data, LEA and school program participation and LEA financial information).

In the future, the ISBE team expects its new system to capture and track longitudinal data on students in Illinois schools, from pre-kindergarten to their employment outcomes. With a longitudinal data system in place, ISBE also plans to improve its ability to support the Federal Electronic Data Exchange Network (EDEN)/EDFacts. ISBE's current system supports a series of automated programs that pull data from various source systems to produce the aggregations and calculations for EDEN/EDFacts. According to ISBE, this is a high maintenance process that can be streamlined with the right data architecture and solution set in place.<sup>39</sup>

### *Lessons Learned*

In March 2010, the ISBE released a Request for Sealed Proposals (RFSP) to contract with a vendor to develop enterprise-wide data architecture. The ISBE plan will include data from 13 different systems that currently use a mixture of Access and SQL servers. These 13 systems range from 150 to 3,000 data elements and use Web, LAN, and standalone applications.

Currently, LDS system is in the design phase. The state recently hired Public Consulting Group to design the data architecture. Their methodology to date has involved interviews with program and technical resources and alignment with data model work they are performing for CCSSO. The ISBE team anticipates the design process will take 6 months and framing the data architecture will take another 6 months.

---

<sup>39</sup> Illinois State Board of Education. (March 2010). Request for Sealed Proposals (RFSP): Data Architecture Vendor for the Illinois Longitudinal Data System (ILDS) Project.

## A.2 North Dakota Department of Public Instruction

State/Agency: North Dakota Department of Public Instruction

Web Site: <http://www.dpi.state.nd.us/>

Address: 600 E. Boulevard Avenue, Dept 201  
Bismarck, North Dakota 58505

POC: Tracy Korsmo

POC Phone: 701-328-4134

POC Email: [tkorsmo@nd.gov](mailto:tkorsmo@nd.gov)

### Case Profile

**Student Enrollment:** 94,728<sup>40</sup>

**Teachers:** 8,181<sup>41</sup>

**LDS Grant:** \$6,723,090<sup>42</sup>

### Background

Prior to 2007, North Dakota did not have a Longitudinal Data System; however, state leaders realized the importance and benefits of linking data among North Dakota Department of Public Instruction (K-12 schools), the North Dakota Department of Commerce, Workforce Division and the North Dakota State Board of Higher Education and pushed for an LDS project. To achieve this goal, these leaders realized that foundational components were necessary and, thus, hired Claraview to develop a state-wide LDS strategic roadmap. Shortly after the roadmap project began, North Dakota applied for and received a Statewide Longitudinal Data Systems grant from the U.S. Department of Education. Future funding for the ND LDS will be a combination of these federal funds and state appropriations.

### System Design and Architecture

Currently, North Dakota's LDS is still in the design stage. The North Dakota LDS team plans to build the K-12 data and Workforce Department warehouse, while the higher education data warehouse will be expanded by the higher education community's IT staff. The three separate systems currently are not integrated, but will be in the future, possibly in 2011.

<sup>40</sup> National Center for Education Statistics. "State Profiles Home Page." National Center for Education Statistics. U.S. Department of Education, Fall 2009. Web. 15 Dec. 2010. <http://nces.ed.gov/programs/stateprofiles/sresult.asp?mode=short&sl=17&s2=38>

<sup>41</sup> Ibid.

<sup>42</sup> National Center for Education Statistics. "Statewide Longitudinal Data Systems Grant Program - Grantee State – North Dakota." National Center for Education Statistics (NCES). U.S. Department of Education, May 2010. Web. 15 Dec. 2011. <http://nces.ed.gov/programs/slids/state.asp?stateabbr=ND>

During the planning stages, the NDLDS team inventoried their existing statewide data sources and discovered that the North Dakota Department of Public Instruction and had their own data warehouses. Therefore, rather than creating an entirely new system, it was a natural match to build out the K-12 warehouse to an LDS system. This approach saved the team time and resources.

North Dakota's K-12, workforce (from the Human Resources Department), and higher education data is expected to have its own separate warehouse. Since K-12 and workforce data are already being stored in a current warehouse, North Dakota is working on expanding and matching these warehouses to a longitudinal data system that is based on the K-12 data warehouse. The SLDS, which is being built as an extension of the K-12 warehouse is to consume data from all the different warehouses. The architecture has not been finalized – issues of whether or not it will be separated from all the other warehouses, security, are still being researched and deliberated. Additionally, how and to what extent the data will be shared outside of internal researchers have not yet been decided. North Dakota may possibly use a separate warehouse and portal for aggregated data to be available for non-state-agency users.

According to North Dakota, its Higher Education agency is the most disjointed part of the LDS – because the agency is building its own and has their own staff. The linkages between the agencies have not yet been started; however the state is now developing a process to align student data. Attributes for the linkages being considered are: name, date of birth, gender, graduating high school, and social security number.

North Dakota has data sharing contracts between: K-12 and Higher Ed; K-12 and Unemployment Insurance; and Higher Ed and Unemployment Insurance. Another issue they are currently working to overcome is governance. Currently, part of the statewide LDS is governed by the state's own privacy and sharing rules. Subcommittees are currently reviewing their privacy and sharing laws and trying to decide if the current legislations are conducive to a properly functioning and streamlined process for an LDS.

Currently, North Dakota does not have an LDS design yet. The state is in the process of finalizing the K-12 project plan. By the end of December, North Dakota hopes to have a project plan in place for its LDS. North Dakota plans to utilize canned reports to share with the public; they do not plan on having ad-hoc reporting capabilities in the near future. In terms of funding, the state is currently looking for ways to augment their current funding source. North Dakota is hoping to be able to secure funding to keep its LDS running once it is built. North Dakota's immediate next steps are to work on a proof of concept, implement its K-12 warehouse, and work on an "entity resolution accountability plan".

### *Lessons Learned*

Because the North Dakota LDS is in the early stages of planning and designing, Ms. Korsmo was unable to provide many recommendations for best practices. However, data governance has emerged as an important issue. Although Ms. Korsmo did not share specifics related to data governance issues, she indicated that the team is facilitating negotiations among the different agencies and current legislation is being reviewed. Ms Korsmo concluded with a recommendation that prior to planning the design of a system, to determine whether there are

existing centralized statewide agency systems upon which to build an LDS. The North Dakota team is optimistic for its plans to expand the K-12 data warehouse and expects this leveraging of the legacy system to reduce project time and costs.

### A.3 Washington Education Research and Data Center

State/Agency: State of Washington Education Research and Data Center

Web Site: <http://www.erd.c.wa.gov/>

Address: 210 11<sup>th</sup> Avenue SW, Room 318

P.O. Box 43113

Olympia, Washington 98504

POC: Dr. Michael Gass

POC Phone: 360-902-0599

POC Email: [Michael.Gass@OFM.WA.GOV](mailto:Michael.Gass@OFM.WA.GOV)

#### Case Profile

**Student Enrollment: 1,037,018**

**Teachers: 54,428**

**LDS Grant: \$17,341,871**

#### Background

The Washington Research and Data Center, an agency under Washington’s Office of Financial Management, manages research and education data for the state and is leading the state’s LDS project. The Center also manages four-year higher education enrollment data systems. The SLDS grant Washington received was based on an inter-agency proposal to support federal projects that included a P-20 data warehouse. Since the beginning of the project in July 2010, Washington has established an executive sponsorship steering committee that currently is working through basic governance issues, defining scope and the project management framework, etc.

#### System Design and Architecture

The SLDS project is focused mostly on an internal ability to mine the data. The proposal submitted to NCES specified the LDS type and what different agencies would be linked. The SLDS team envisions a P-20 data warehouse that would receive data from the K-12 warehouse that currently is being built. Additionally, the LDS will “inherit the model” K-12 builds, though designs on the integration model are not yet in place. Washington’s LDS also will integrate data from its stakeholders:

- Office of the Superintendent of Public Instruction (OSPI)– K-12 data
- Department of Early Learning
- Higher Education Coordinating Board
- State Board of Education
- State Board for Technical and Community Colleges
- Council of Presidents – a voluntary association of the presidents of Washington State's six public baccalaureate degree granting institutions
- Employment Security Department – provides employment and employer data
- Professional Educator Standards Board

- Higher Education Coordinating Board
- Independent Colleges of Washington
- Workforce Training and Education Coordinating Board

Since OSPI contracted out its K-12 data system that may influence the design of the larger LDS. OSPI contracted with a vendor to customize its existing longitudinal system. The system will be built over the next year.

The Washington State team also is investigating the initiative with the CCSSO (Council of Chief State School Officers) model, which is a P-20 core modeling project to be designed by Public Consulting Group (PCG) and funded by the Gates Foundation. The project involves building a data dictionary and longitudinal data model. Several states currently are participating in this initiative and may use the resulting system. The Washington SLDS team is considering using this model if the state will build internally and will not purchase an off-the-shelf product.

The Washington LDS team has yet to determine its LDS' technical architecture. However, the state will be using an SQL server – as it is already using it in current systems. The greatest issue is whether or not its LDS will be outsourced because the state will have less control of its system if it is outsourced. Additionally, Washington is considering a complete Microsoft BI package such as SharePoint.

### *Data Usage and Reporting*

The vision for the system was a normalized data set subscription service at least through an authenticated site, which would provide a number of reports. Currently, the team anticipates that matches among agencies will be primarily through social security numbers (SSN), although the state will use a centralized matching system that relies on more than social security numbers. High school and post-secondary institutions will be matched through social security numbers, but K-12 data will be matched differently. Reports will be provided for both aimed and anonymous users.

### *Lessons Learned*

Currently, Washington has been in a state of political impasse. The various agencies have not been able to agree on an IT governance model, although Dr. Gass declined to provide specifics. It is an issue the team hopes to overcome within the next few months, especially since there is still much to be done and funding from the federal government will end in July 2013. The LDS team plans are to complete the system design in one year and two years of active development.

## **Appendix B: Best Practice Case Studies Interviewee List**

### ***B.1 Indiana Department of Education***

**Contact:**

Molly Chamberlin  
Director, Data Analysis Collection and Reporting  
Indiana Department of Education

**Email/Telephone/Address:**

[mchamber@doe.in.gov](mailto:mchamber@doe.in.gov)  
317-234-6849  
151 West Ohio Street  
Indianapolis, Indiana 46204

**Links:**

- <http://www.doe.in.gov/data/> - Indiana Department of Education Data Web site

### ***B.2 Iowa Department of Education***

**Contact:**

Jay Pennington  
Bureau Chief  
Iowa Department of Education

**Email/Telephone/Address:**

[Jay.pennington@owa.gov](mailto:Jay.pennington@owa.gov)  
515-281-4837  
400 E. 14<sup>th</sup> Street  
Des Moines, Iowa 50319

**Links:**

- [http://www.iowa.gov/educate/index.php?option=com\\_content&task=view&id=1691&Itemid=2490](http://www.iowa.gov/educate/index.php?option=com_content&task=view&id=1691&Itemid=2490) – EdInsight Web site

### ***B.3 Data Strategies – Army Suicide Mitigation Project***

**Contact:**

Susan Carter  
Managing Partner  
Data Strategies  
Kevin Corbett  
Managing Partner  
Data Strategies

**Email/Telephone/Address:**

[SCarter@DataStrategiesInc.com](mailto:SCarter@DataStrategiesInc.com)

814-965-0003  
[KCorbett@DataStrategiesInc.com](mailto:KCorbett@DataStrategiesInc.com)  
P.O. Box 772  
Midlothian, Virginia 23113

## ***B.4 Texas Education Agency***

### **Contact:**

Brian Rawson  
Director of Statewide Data Initiatives  
Texas Education Agency

Nina Taylor  
Director of Information Analysis  
Texas Education Agency

### **Email/Telephone/Address:**

[brian.rawson@tea.state.tx.us](mailto:brian.rawson@tea.state.tx.us)  
513-936-2383  
[Nina.taylor@tea.state.tx.us](mailto:Nina.taylor@tea.state.tx.us)  
512-475-2085  
1701 North Congress Avenue  
Austin, Texas 78701

### **Links:**

- <http://www.tea.state.tx.us/> - Texas Education Agency Web site
- [http://www.texaseducationinfo.org/tpeir/TPEIR\\_Documentation.pdf](http://www.texaseducationinfo.org/tpeir/TPEIR_Documentation.pdf) - Texas Public Education Information Resource

## ***B.5 Data Strategies – DLA Data Convergence and Quality Project***

### **Contact:**

Susan Carter  
Managing Partner  
Data Strategies

### **Email/Telephone/Address:**

[SCarter@DataStrategiesInc.com](mailto:SCarter@DataStrategiesInc.com)  
814-965-0003  
P.O. Box 772  
Midlothian, Virginia 23113

## ***B.6 NORC Data Enclave***

### **Contact:**

Timothy Mulcahy  
Senior Research Scientist

## NORC at the University of Chicago

### Email/Telephone/Address:

[Mulcahy-Tim@norc.org](mailto:Mulcahy-Tim@norc.org)

301-634-9330

University of Chicago

4350 East West Highway

Bethesda, Maryland 20814

### Links:

- <http://www.norc.org/DataEnclave> - NORC Data Enclave Web site

## ***B.7 Illinois State Board of Education***

### Contact:

Michael McKindles

ILDS Project Manager

Illinois State Board of Education

### Email/Telephone/Address:

[mmckindl@isbe.net](mailto:mmckindl@isbe.net)

217-782-0329

100 N. 1st Street

Springfield, Illinois 62777

### Links:

- <http://www.isbe.state.il.us/ILDS/htmls/project.htm> - Illinois Longitudinal Data System Project Web site

## ***B.8 North Dakota Department of Public Instruction***

### Contact:

Tracy Korsmo

Business Intelligence Program Manager

North Dakota Department of Public Instruction

### Email/Telephone/Address:

[tkorsmo@nd.gov](mailto:tkorsmo@nd.gov)

701-328-4134

600 E. Boulevard Ave., Dept. 112

Bismarck, North Dakota 58505

### Links:

- <http://www.dpi.state.nd.us/> - North Dakota Department of Public Instruction
- <http://www.nd.gov/itd/planning/initiatives/roadmap.pdf> - State of North Dakota Longitudinal Data System Strategic Roadmap

## ***B.9 State of Washington Education Research & Data Center***

**Contact:**

Michael Gass  
State of Washington Education Research & Data Center

**Email/Telephone/Address:**

[Michael.Gass@ofm.wa.gov](mailto:Michael.Gass@ofm.wa.gov)

360-902-0599

210 11<sup>th</sup> Avenue SW, Room 318, P.O. Box 43113

Olympia, Washington 98504

**Links:**

- <http://www.erd.c.wa.gov/default.asp> - Education Research & Data Center website
- <http://nces.ed.gov/programs/slds/pdf/washingtonabstract2009ARRA.pdf> - Project Abstract from Department of Education website.

## Appendix C: Materials Sent to Best Practices Interviewees

### C.1 Best Practices Interview Template

#### General Project Overview

1. What were the objectives of the project?
2. Do you have a project abstract/overview that you are able to share?
3. Who were the stakeholders?
4. How long did the project take?
5. How much did the project cost?

#### Database Design/Architecture

6. What steps were involved in designing the database/warehouse?
7. What architecture existed within your organization prior to the design and implementation of the system?
8. Describe the general architecture of your system and its different components. Is there a model that you use (federated, non-federated)?
9. Do you have a visual representation of the system that you are able to share?
10. What products are used for the underlying data management (e.g. DBMS)?

#### Data

11. How much data flows through the system (e.g. number of records)?
12. Were there disparate data source?
13. Was sensitive (PII) data contained within the source database?

#### Security

14. What were the security requirements of this system?
15. Does your system de-identify personal data? If so, please describe your data-de-identification process.
16. Does the system contain an authentication process? Is it single or dual factor

#### Users

17. Can you describe the different users of the system?
18. Were separate processes (databases) used for anonymous users and named users?
19. What level of help desk support was provided?

#### Implementation

20. Can you give me a picture of what was involved in the implementation process?
21. What was the level of effort? How many man-years did the implementation take from design to implementation?
22. What were the barriers/challenges of implementing this program?
23. What ongoing efforts and resources are need after a system is up and running?
24. Knowing what you know now, how would you approach the problem and implementation differently?

#### Performance & Feedback

25. How is performance of the system measured?
26. How is performance affected through variations within the system?
27. Were any compromises made to the system design to in order to achieve acceptable levels of performance?
28. Describe the results of the system and its ongoing use.

### Other

29. What additional best practices or lessons learned can you share?
30. Are there additional resources (e.g., people, documents, links, etc.) that would be helpful?
31. What are the next steps?

## **C.2 Architectural Best Practice, Design & Planning Support Project Overview**

The LDS will be a logical data warehouse that is fed by numerous agencies from around the Commonwealth. The data provided by each agency is a subset of the agency's primary data repository (database). It is possible that more than one data source from an agency may be required. The linkage between the agencies is focused on a student as s/he progress through the school system into the workforce. LDS-level information will be de-identified so that no individual can be uniquely identified from the data. Even though the linkage between the agencies is student-centric, the management of the data in the LDS is an unidentifiable person. It is imperative that the design of the LDS protect the individual and not fall subject to re-identification.<sup>43</sup>

External query interfaces into the LDS will be made available to the public, partner agencies, and researchers. Level of access to the LDS data is managed through an authentication and authorization scheme. Levels may be from selection of pre-canned queries to ad hoc requests.

The objective will be to develop a high-level architecture and development strategy which can then be used to determine sourcing requirements, staffing levels and become the input to a detailed design and staffing phase which will follow rapidly from this project. General design questions regarding the project include:

- What components of the solution need to be developed (and what needs to be procured)?
- What are the key development tasks?
- What actual processing and response speeds are required and what is acceptable?
- What will be best way to de-identify the data? What is the best way to link de-identified data?

CIT will perform research and analysis of other data warehouse designs, integration efforts, and their implementations from similar government and corporate organizations. From a series of research, interviews, and surveys, CIT hopes to develop best practices and lessons learned from such programs. Participation in CIT's research will help the Virginia Department of Education in deciding what plan of action to execute.

---

<sup>43</sup> Ochoa et al., *Reidentification of Individuals in Chicago's Homicide Database: A Technical and Legal Study*, tech. report, Massachusetts Inst. Of Technology, 2001. Retrieved from:  
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.15.7467&rep=rep1&type=pdf>

## Appendix D: Materials Sent to Subject Matter Experts

Prior to an interview, each Subject Matter Expert was provided a package of pre-reading material that helps familiarize him/her with the SLDS program. The package consisted of a sample set of questions that may be asked during the interview and a white paper on the grant award, the conceptual architecture, constraints placed on the implementation, and sample use cases for the system.

### D.1 Subject Matter Expert – Interview Template

#### *Data Modeling*

1. What do you consider the key steps to a successful integrated data model when incorporating numerous data sources?
2. Can you describe the level of complexity for a data warehouse data model versus a federated data model?
3. What is a common problem overlooked in data modeling when dealing with multiple data sources?

#### *Security (General)*

4. In the area of data governance with numerous data owners, are there techniques that make data access controls easy to manage and implement across data sources?
5. Have you been involved with a data breach and associated legal issues that occurred? Do you have any recommendations on how to best avoid a data breach?
6. Are there enhanced security models that are used with database management systems that add an additional layer of protection (e.g. additional firewall in front of DBMS server)?
7. When implementing an in-memory database (IMDB), are there security/protection techniques that differ from a standard disk-based data warehouse/federated?

#### *Security (De-Identification)*

8. Can you describe different techniques/algorithms used for data de-identification?
9. Are you familiar with any known issues in de-identification algorithms that could cause data elements to be re-identified?
10. What issues existing in attempting to update de-identified data from its original source? Are there good techniques for achieving this process?
11. Can the de-identification process add substantial processing time to a database transaction/query? If so, are there techniques to help minimize the processing overhead?

#### *Architecture Trade-offs*

10. Given the described conceptual design, what different implementation models might best work and why?
11. For any of the given models, are their known constraints/limitations on the implementation (e.g. one model works best for small data sets)?
12. For a federated model, what can impact the performance for data access?
13. If the database is designed for online analytical processing, does that have an influence on the type of implementation model?

14. Are there best practices for linking data records across a federated data model (when a common unique ID is not present)?
15. Can you describe different techniques/algorithms used for linking data across a federated model?
16. Are in memory databases trending up or down and why?
17. What lessons learned are there with using an in memory database?
18. Are there advantages or disadvantages to using an in memory database to replace a standard data warehouse?

### *Query Processing*

19. Are there preferred query strategies for optimization when the system is implemented as a singular data warehouse versus federated?
20. In a federated model, is the query optimization better handled in centralized, distributed, or hybrid implementation?
21. What impacts from the network topology can affect the query processing implementation?

## ***D.2 Virginia Statewide Longitudinal Data System - Executive Summary***

### ***Collaborative Partnership***

The Statewide Longitudinal Data System (SLDS) is a collaborative effort by the Virginia Department of Education (VDOE), Virginia Employment Commission (VEC), State Council of Higher Education for Virginia (SCHEV), Virginia Community College System (VCCS) and Virginia Information Technologies Agency (VITA).

The Commonwealth of Virginia's Department of Education (VDOE) successfully secured a multi-year federal grant for the design, development and operation of a Statewide Longitudinal Data System (SLDS) to integrate student and workforce data in the Commonwealth. Specifically, the SLDS will integrate K-12, higher education and workforce data into a single logical database which can be used for research and analysis. Elements of the SLDS project will focus on transactional data (e.g., transcripts and student records), which will reduce the cost burden for a number of education stakeholders, including students, parents, administrators, school counselors, registrars, and college admissions officers. Other elements will focus on the integration and delivery of de-identified data via a web portal, and the management and governance of the Commonwealth's education and workforce data.

In order to establish this comprehensive, longitudinal data system, the SLDS will be developed in phases, with the initial phase creating a federated longitudinal data linking and reporting system linking data among state agency data sources, including K-12, higher education, and workforce systems. A rubric will be created to document data element definitions, data requirements, and technical requirements for de-identified data sets that can be linked among agencies; build a central linking directory based on data sharing agreements in place or established as part of the grant project; and establish a query process for authorized user access that uses the linking directory to anonymously join individual-level records from multiple data sources.

### ***Stakeholders***

The SLDS will serve a variety of stakeholders, to include legislators, policy makers, teachers, school administrators, education program directors, researchers (both inside and outside the state), parents, localities, citizens and the media. Appendix A provides a simplistic view of potential uses by stakeholders.

### ***Future Benefits***

When complete, the SLDS will provide researchers, analyst, educators, parents, students, policy makers, and program administrators with the following business benefits:

- Establishing kindergarten to college and career data systems that track progress and foster continuous improvement;
- Enhancing the Commonwealth's ability to examine student progress and outcomes over time by linking individual-level student data from K-12 education, postsecondary education and the workforce system;
- Enabling the exchange of data among agencies and institutions within the State and between States to inform policy and practice;
- Linking student data with teachers primarily responsible for providing instruction;

- Enabling the matching of teachers with information about their certification and teacher preparation programs;
- Enabling data to be easily generated for continuous improvement and decision-making
- Ensuring the quality and integrity of data contained in the system;
- Enhancing the Commonwealth’s ability to meet reporting requirements of the U.S. Department of Education;

### *Technology Challenge*

Virginia stands out as a special case study in the difficulties of combining data from multiple agencies. In addition to the standard layers of complexity, Virginia-specific privacy laws and historical system of “locally administered, state supervised” public services create additional challenges. This complex network of technological, regulatory, and structural impediments to the integration of individual-level data makes a **traditional approach—consolidation of data in a physical central ‘warehouse’—untenable**. To successfully combine Virginia’s set of heterogeneous data sources; Virginia proposed a federated data systems approach.

### *Federated Data Systems*

Virginia’s federated data system will interact with multiple data sources on the back-end and present itself as a single data source on the front-end. The key to successfully linking the different data sources is a central linking apparatus. Generally, this is a database management system that has been set up with access privileges to each data source and that houses a ‘linking-table’ populated with the unique identifiers that will be used to ‘join’ the tables together into one large data set.

### *Using a federated system to merge data across agencies*

The special requirements imposed on the establishment of a federated data system between public agencies dictate that individual privacy be maintained. For example,

While it may be acceptable for any system user to retrieve information showing that a certain individual participated in a certain program (e.g., William Smith went to Virginia Tech), it would not be acceptable if that same user could link that person to a particular detail specific to that program (e.g. William Smith received a grade of “F” in Calculus). However, for the purposes of longitudinal research, it is that latter information that is needed (e.g., we don’t care about William Smith, but we do want to know how many males failed Calculus in a particular year).

Therefore, what is needed is a system that will permit the linking of data relevant to longitudinal research that does not allow personal identification of any of the individuals in the data set. The proposed solution has two distinct processes, one for establishing and maintaining an anonymous ‘linking directory,’ and one that uses the linking directory to join data sources and return a ‘de-identified’ data set to the user through a data query process.

Virginia has proposed a cross-agency data linking and reporting system that can be used in a manner that maintains the confidentiality of individual student/teacher/employee data, can be used for accountability and analytic purposes, and meets the requirements of State and federal privacy laws. In Virginia, state law (§2.2-3800- § 2.2-3816) currently prohibits state agencies from sharing personal information across state agencies except under specific circumstances.

In order to meet SLDS program requirements and the needs of the collaborating state agencies, Virginia proposed a methodology that would permit multiple state agencies to merge de-identified individual-level data using a federated data system model. The methodology, developed

in conjunction with Virginia’s Office of the Attorney General to link data between K-12 and higher education, will permit the two education systems and the many agencies that house workforce education and training programs to link unit-level records through the use of de-identified data sets.

### *An Example*

Suppose a request is received for a report on college students who entered Virginia’s community college system for the first time in 2006, and the variables of interest are participation and outcomes on statewide assessments in high school; credential information as of August 31, 2009; and employment outcomes since the student left high school.

Using data element definitions in the system, the user will define the cohort and variables to include in the dataset. Then, the system will identify students for study in the central directory, and, using the hashed identifier in that directory, join data from participating agency systems. Finally, joining tools will replace the identifier in the central directory with another unique hash value for each individual in the dataset, and deliver the final de-identified data in the format specified by the user.

### *Summary*

The objective of the SLDS is to propel Virginia’s data collection, reporting, and analytic capabilities far beyond current capacities by merging K-12, higher education and workforce data. By merging de-identified data in a federated system, Virginia will maintain compliancy with state and federal privacy laws, while meeting critical data reporting requirements and policy-development needs.

## ***D.3 Virginia Statewide Longitudinal Data System - Usage***

### ***Actors:***

Virginia's SLDS will be utilized by a variety of stakeholders, to include legislators, policy makers, teachers, school administrators, education program directors, researchers (both inside and outside the state), parents, students, localities, citizens and the media.

### ***Potential Use:***

The following statements represent potential use of the system by actors:

- Actors will use a web-based portal to access publicly available data from the K-12, postsecondary education, and workforce agencies.
- Actors will access/create reports that will be available in a variety of formats depending on the user's preferences. Tables, charts, and graphs will be presented to provide different views of the data. Maps will also be used to provide a geographic perspective. GIS data layers, including county and city boundaries, roads, schools, school districts, and related information such as census counts and income levels will be integrated with the geographic reports to provide contextual information for the data and for further analysis.
- Actors will develop custom reports by combining data from multiple publicly available datasets, and be able to request the data in multiple formats.
- Actors will develop Custom reports by identifying the cohort, independent and dependent variables and selecting an output method (table, chart, graph, Excel, CSV, etc.)
- Actors will view [create] a report identifying the number and percentage of teachers and principals rated at each performance rating or level.
- Actors will create [view] a report showing growth data for current and previous year students and estimates of teacher impact on student achievement.
- Actors will view a report of high school graduates who enroll in state institutions of higher education and complete at least one year's worth of college credit within two years.
- The proposed information system will also include reporting capabilities available to teachers and other authorized school division personnel to provide estimates of student growth and teacher impact on student performance on state assessments in reading and mathematics.
- Actors will create reports that link students to course enrollment, course grades, and to the teachers providing instruction in each course.
- Actors will be able to view pre-developed and publicly available reports created by VDOE, SCHEV and VEC.
- Actors will use the longitudinal data system to complete portions of reports required by lawmakers, such as a 2007 study of high school dropout and graduation rates
- Actors will receive information about students who have failed state-wide assessments for two or more years in a row

- Actors will use the SCHEV Student Data Warehouse to create standard and ad hoc reports on postsecondary education.
- Actors will view standard reports that are publicly available on VDOE's website, including summary data for required and commonly requested information such as numbers of students enrolled, and graduated, dropped out, and participating in special educational instructional services.
- Actors will develop additional reports using data to be collected for the student-teacher information system, combined with information already collected at the student level, to develop reports that provide comparison of end-of-course grades with performance on state assessments, and additional information on students not tested by grade and subject.
- Actors will conduct analyses of specific content standards that, when met, describe the type of work that students must achieve to be ready for postsecondary education.