

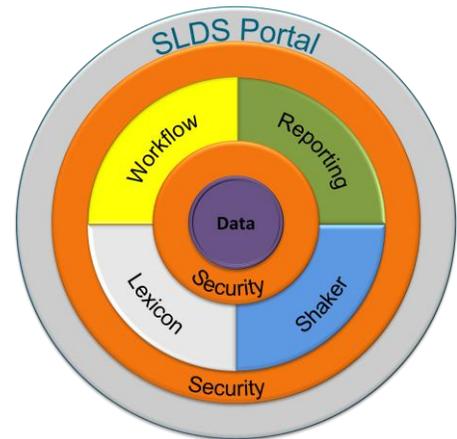
# State Longitudinal Data System Architecture Review Summary

---

## Overview

On January 27, 2011 the VITA and Virginia Tech team presented the logical architecture and associated physical infrastructure for Outcomes 2 and 3 of the Statewide Longitudinal Data System (SLDS) to the Commonwealth of Virginia stakeholders. The two hour meeting covered the seven components that comprise the architecture, recommended commercial products where applicable, and provided insight into the use of shared resources in the physical implementation.

The SLDS architecture was represented by a bull’s-eye signifying the data-centric nature of the architecture. The importance of security was 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.



## SLDS Components

At its highest level, the SLDS system provides users with access to data. Through the SLDS Portal, the public will have access to data released by participating agencies for public use. Users interested in accessing additional, non-public data will access the SLDS Portal to request access. Once this request is approved by authorized agency staff, the user will log on to formulate a request for data, in the form of a query. Once approved, the request will be executed by the SLDS Shaker, and the results returned to the user via secure mechanisms.

The SLDS architecture is comprised of seven functional components that work together to deliver these requirements. Requirements will be developed in detail, and the components are only summarized here.

Each component was presented in detail during the review. During the presentations, recommendations for commercial products were raised where appropriate, and clarifications for requirements surfaced. A summary of each component presentation follows.

## Portal



A mock-up of the SLDS Portal illustrated the numerous functions the Portal provides for both public (anonymous) users and named users (authorized agency personnel and researchers). The front door into the SLDS is through the Portal, regardless of the user type.

The portal serves both public and named users. Named users gain access after they have requested an account, and their request has been approved by the necessary agencies. The approval process will incorporate procedures required to meet state and federal legal and regulatory requirements for access to restricted data. The named-user path provides access to help, training, the components of the Lexicon, data requests and status of requests, and account maintenance including password reset.

The public-user path through the Portal provides access to help files, frequently asked questions, aggregated data reports in multiple formats, data definitions and descriptions of summary data, and links to agency specific reports. The amount of metadata viewable by the anonymous user is still to be defined.

## Security



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

*Authentication*, or verification of the identity of a user, is required for all named users, to include agency employees and agency-approved researchers. Researchers and agency employees will be authenticated as a precondition to gaining access to the named-user portions of the SLDS portal. Additionally, agency employees will be authenticated before gaining access to the Workflow component of the SLDS application.

In order for an agency employee to interact with the Workflow component, s/he would need to log in (authenticate) to the Commonwealth of Virginia (COV) infrastructure. Thereafter, s/he would be able to access the Workflow component in order to act on the Workflow trigger.

*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 (permissions) the Lexicon. The Workflow component is the hub for managing the users' roles and associated permissions. The SLDS components coordinate with the Workflow to manage requests for services correctly.

## Workflow



The Workflow component manages the account creation and data request functions. Workflow monitors and triggers actions such as query submission and also maintains status of requests. Moreover, Workflow is the source of information about roles and permissions for SLDS users.

*Account Creation.* When a named user requests creation of an SLDS account, it is the Workflow component that manages the message(s) and any required documentation, and notifies designated COV employees about the request. Through the Workflow component, employees can review the request and supporting documentation required to meet legal and regulatory requirements, and approve or deny the request. Workflow then notifies the submitter of the account request of the final decision.

*Data Request.* When a named user makes a data request the Workflow component monitors the request, confirms approval, and submits the query to the Shaker for the action. Designated COV employees are notified of the request to approve or deny the query request. If approved, Workflow then submits the request to the distributed query engine and continues to monitor the status. Upon completion of the transaction, Workflow would notify the researcher the data set is available for download via secure mechanisms.

## Reporting



The Reporting component provides two functions: aggregate reports and record-level reports. The aggregate report is designed for the public (anonymous user). The record-level report is designed for the agency employee or agency-approved researcher (named user) who has been granted permission to view a finer resolution of data.

The aggregate reports will be prebuilt reports and based on commonly requested or high interest data. The interface to these reports may provide the user with limited capabilities to perform analysis on the data (e.g., grouping, filtering, etc.). Periodically, the prebuilt aggregate reports will be updated through a background “extract, transform, and load” (ETL) process.

The record-level reports are the end result of a named user building a query using the Query Building Tool (QBT). The query can be tested against a shell database that contains mock data. This will allow the named user to evaluate the query and mock query results before the query is submitted to the applicable agencies for approval. Since the process is typically not time sensitive, the named user will be informed that s/he will receive a message when the report (data set) is ready.

When the submitted query is completed, the resulting data are deposited in a user repository. The named user is notified of the completion, and informed via the Workflow component the data is available for downloading via secure mechanism. The user can download the file, and display the data with a simple analysis tool.

## 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.”*

A named user, when building a query, will interact with a set of field names and relationships to formulate a query. The user interface for the query building will provide a simple view of the relevant components of the Lexicon, such as data definitions, possible linking methods, and descriptions of limitations of what can and cannot be linked reliably. This information would be designed to provide for easy query construction.

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

- Changes in data ranges
- New data fields
- 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.

## Shaker



The Shaker is responsible for querying the agency data sources, matching the records in a de-identified manner, creating final de-identified data sets, and communication with the workflow system. To successfully execute these responsibilities, the Shaker must have an understanding of the structure of the data sources and the relationships between them. The Shaker will use the Lexicon to obtain this structure and relationship information.

In querying the agency data sources, the Shaker has responsibility for dividing the submitted query into smaller queries (sub-queries) and devising an optimized plan of execution that can be executed against each target data source. Depending on the form of the query (e.g., inner join, left join, equijoin) and the specified final output parameters (e.g., counts of non-matching records by demographic categories), a different execution plan will be employed.

To enable de-identified matching of the records generated from the sub-queries, the Shaker has responsibility for the operation of an anonymous linking process based on hashed-keys. In this process, the Shaker generates a “key” for each submitted query that is then used by each data source to generate a secure one-way hash of uniquely identifying information for each individual.

Therefore, in the data source’s response to the sub-query submitted by the Shaker, the newly generated one-way hash is sent in place of the uniquely identifying information.

In creating the final de-identified data set, the Shaker has responsibility for collecting the sub-query results and joining them together using the one-way hash to produce a set of joined data. Upon successful record joining, the one-way hash is deleted from the data set and replaced with a random ID which cannot be traced back to the original data source(s). Upon successful creation of the de-identified data set, Workflow is notified of the status and location of the file. The query submitter (e.g., researcher) is then notified that the data is available.

**Data**



In addition to the primary data sources in the SLDS, there are numerous small databases residing in the system. Even though the end product, i.e., reports, is driven by the primary data sources, the supporting databases are critical to the implementation of the SLDS. These supporting databases support data requirements for the Query Building Tool (QBT), Shaker, logging, auditing, reporting, Workflow, etc.

The Shaker process will use its database to temporarily store linked record-level data, to create pre-built, public-facing, aggregated-data reports. The temporary tables may be archived if this is determined to be a requirement, but the temporary tables will not be stored in a manner that makes them readily accessible for queries or reports. For pre-built reports, stored procedures in a database will be used for data querying and suppression.

**Physical Infrastructure**

Three physical infrastructures were proposed. The first infrastructure involves the composition of the production environment, which would be located at the CESC. Many of the servers shown would support multiple functions for the SLDS. This approach as well as the use of existing shared services resources helps to keep costs low.

The second infrastructure involves the test environment, which would also be located at the CESC. Once again, servers would be multi-functional, and existing servers would be incorporated as appropriate. Target data sources in the test environment will be coordinated with the stakeholders.

The final infrastructure involves the development environment. This will be a hosted environment managed outside of the Commonwealth of Virginia’s infrastructure. This configuration provides the greatest flexibility for the development of the SLDS. As in the test environment, data for the simulated data sources will be coordinated with the stakeholders.

### COTS vs. Custom

Throughout the presentation, the team presented options for COTS and custom solutions. The following table summarizes the recommendations and positions of the team.

Component		Custom / COTS	Suggested Product
Portal		Custom	
Security		Custom	
	Authentication		COVAUTH
	Authorization	Mixed	Mixed
Workflow		COTS	MS Dynamics
Reports			
	Public Facing	COTS	Logi Info
	Query Building	COTS	Logi Ad-Hoc
Lexicon		Custom	
Shaker			
	Extract, Transform & Load	COTS	Logi ETL, SSIS or Informatica
	Distributed Query Engine (DQE)	Custom or COTS	Syncsort, Informatica or Custom