IES Institute of Education Sciences

# Historical Data Migration

## SLDS Guide

Statewide Longitudinal Data Systems Grant Program

November 2021

U.S. DEPARTMENT OF EDUCATION

*A Publication of the National Center for Education Statistics at IES*

## Contents



This product of the Institute of Education Sciences (IES) Statewide Longitudinal Data Systems (SLDS) Grant Program was developed with the help of knowledgeable staff from state education agencies and partner organizations. The information presented does not necessarily represent the opinions of the IES SLDS Grant Program.

For more information on the IES SLDS Grant Program or for support with system development, please visit *http://nces.ed.gov/programs/SLDS*.

**CONTRIBUTORS**
Duane Brown and Trcica Farris, *CEDS Support Team*
Kathy Gosa, *SLDS Grant Program State Support Team*
Drew Bennet and Wendy Geller, *Vermont Agency of Education*
Heather Handley and Michael McGroarty, *Michigan Center for Educational Performance and Information*
Rick Thompson, *South Carolina Department of Education*
Data Standards-Based Infrastructure Implementation Workgroup members

## Introduction

Many education data systems have been in place for a number of years and have accumulated longitudinal data that allow states and organizations to follow cohorts of students from prekindergarten through high school, and even through postsecondary education and beyond. However, as states are modernizing their statewide longitudinal data system (SLDS) infrastructure to take advantage of newer technology, streamlined processes, and standards, they must consider what to do about the vast amount of data collected and stored within older technologies and structures.

## Historical Data Migration Considerations

States may choose to migrate all, some, or none of the historical data they have accumulated over the years. There are several factors to consider when determining how to handle historical data:

- **The impact of moving to standards (e.g., data standards definitions and code sets).** For example, in some cases, the state's historical data may include situations where the definition of a data element and its option set codes actually are a combination two or more elements. When moved, will there be enough information to convert this data to the standard?
- **Known data quality issues and concerns.** For example, the historical data could include a dataset that was not fully validated during collection. Is there a way to document the uncertainty of the quality of the data in the new system? Is there a risk to combining these data with other datasets that will be used for longitudinal research?
- **Absence of business rules and other documentation for the historical data.** For example, the historical dataset could contain a list of elements, but the rules that governed the collection and meaning of the data for those elements is not available. Is there a need to move or convert the data without knowledge of the true meaning? Will including the historical data introduce risks when using the data? Is there a way to document the uncertainty in the new data model for this specific set of data?
- **Whether historical data are unit level, aggregate, or a combination.** For example, a dataset could include elements that are only aggregate values, whereas the new system may capture the unit-level elements. It may be difficult (or impossible) to move the aggregates to the new standard. Should each set retain its original data model? Will there be a need to compare the new data model aggregates with the older data model aggregates? Are they comparable?
- **The use cases for historical data in the new system.** For example, a dataset no longer is collected, is seldom used, and used primarily for research purposes. Should this dataset be migrated given the needed level of effort? Is it worth the time and resources to convert the dataset to the new standards?
- **Storage space requirements (e.g., type, amount, costs).** For example, if a plan calls for "moving everything" to the cloud, what does that mean? Are there some datasets that are rarely accessed and should not be moved or converted due to storage costs, or that could be moved to a less expensive cloud location or storage solution?
- **Technology limitations and skill set requirements.** For example, the historical data may be sitting in an old mainframe. Are there staff available with the skillset needed to migrate the data to the new architecture? Or is migrating the data a high priority because the skills needed to maintain it are becoming impossible to find?

## Vermont's Approach

Historical data play a large role in the Vermont Agency of Education (AOE) and the Vermont Data Management and Analysis Division's (DMAD) future plans for longitudinal data. The agency plans to use historical data to develop trend analysis and modeling. Historical data will help create statistical process controls to verify data quality. AOE also plans to use historical data to engage stakeholders by developing a more robust modern semantic layer that informs users about the Vermont school system historically, currently, and prospectively.

### *Approach and considerations*

AOE's unit-level and aggregate-level data are stored in a Structured Query Langauge (SQL) Server environment. Unit-level historical data in the SQL Server came from a legacy Oracle environment, historical flat files, and Microsoft Access databases. DMAD is migrating from its current data standards and SLDS platform to the Common Education Data Standards (CEDS) model within a SQL environment. Aggregate-level historical data exist in two servers, the Vermont Education Dashboard and Annual Snapshot. Both exist as modern semantic layers so stakeholders can make intuitive decisions. For example, the Annual Snapshot serves as Vermont's means of displaying data related to its Every Student Succeeds Act (ESSA) State Plan and Education Quality Standards. Administrators with role-based access to their organization's data might make decisions about their Continuous Improvement Plans based on their performance as displayed by the indicators in the Annual Snapshot.

Standardization is a key factor in DMAD's migration approach. DMAD chose to migrate to the CEDS model to increase the efficiency of managing and stewarding data as an enterprise. A shared, standardized data model provides efficient data quality and integrity processes, as well as data use for building reporting products. This approach also will help DMAD and the AOE develop longitudinal analyses in the future. Having access to historical data in the longitudinal model makes user queries more effective toward finding the next steps for the AOE and with monitoring data quality.

Resources also have factored into DMAD's migration approach. Staff resources are limited, and DMAD must operate at maximum efficiency. Using CEDS has helped with this effort as standardization prevents DMAD from getting locked in with a single vendor and its formatting. Lack of resources tends to result in technical debt or the implied cost of future rework as a result of choosing an easy solution in the present. To relieve technical debt, DMAD is using a modern data science tool set (Python and SQL) to move toward a more scalable, sustainable, repeatable, monitored, and documented way of stewarding data.

### *Governance*

To develop the extract, transform, load (ETL) process for the CEDS-compliant data model, DMAD staff need to collaborate with subject matter experts to ensure that historical data are landing in the appropriate place within the CEDS data model to populate reporting products. This work includes documenting historical data naming conventions as part of the transformation to the new CEDS-based naming conventions so that these data are moved into the CEDS model accurately. DMAD engages the data users as part of its testing process to make sure this modernization work treats the historical data appropriately. This process requires conversations with team members about roles and responsibilities for data management, as well as documentation of the lifecycle of these data, from collection to reporting.

### *Challenges*

As with other states, standard formatting has proven to be a challenge. Historical and current data were collected in different formats although they live on the same server. Data migration means that DMAD must align two different collection systems before creating an integration package.

## South Carolina's Approach

Migrating historical data solves several problems for the South Carolina Department of Education (SCDE). Currently, the data are stored in period-based batches —snapshots that represent the data at a particular point in time, such as "Schoolyear 2020, 45th day." This method of storage makes temporal comparisons or detecting trends difficult and costly in terms of processor and programmer time. Migrating historical data to a single container reduces the time and effort needed for reporting. Migration also will reduce the space needed to store the data. Batch-based storage often results in massive duplication of records.

### *Approach and considerations*

As SCDE migrates historical data into the CEDS Data Warehouse, it has discovered several issues with historical data that have affected its approach. During the 2019 school year, SCDE began recording the data to be migrated using CEDS elements and structures. SCDE is building the initial data warehouse of longitudinal records with data from 2019 to the present. After that, it will import data prior to 2019 into the SLDS in period-based batches, working backward in time. SCDE's historical records to import date back to school year 2010, and the format and requirements have changed over that time. For pre-2019 records, staff members will have to alter mapping to accommodate earlier data structures while feeding them into the current CEDS structure. To do this, SCDE created a mapping structure and tools based initially on the CEDS Align tool. This mapping utility was expanded to control data import and flow throughout the system. It records metadata on the mapping information akin to source code control to indicate what set of mapping information to use on a particular set of input data. As variations occur in the historical data, the mapping process can accommodate them dynamically using the different versions of the mapping information.

### *Challenges*

SCDE has encountered several challenges while migrating data. Historical layout and data usage and significance changes often are documented poorly. In some cases, knowledge about particular changes is sketchy or resides in the heads of long-term employees, rather than in permanent documentation.

As records are converted into the new data standard, staff members must account for numerous exceptions that may occur. For example, a current student may have transferred from out of state and their transcript records were entered as equivalent courses with their respective final grades. However, those out-of-state records will not be tied to the district's section enrollment data in its student information system (SIS). Likewise, an SIS may not retain certain information past a retention period, so

that grade records from 4 years ago may not have the accompanying section enrollment information.

## Michigan's Approach

Michigan's Center for Educational Performance and Information (CEPI), located in the State Budget Office, manages a centralized SLDS that draws data from many state agencies; public community colleges and four-year universities; state health, treasury, workforce, police, and public safety programs; and the National Student Clearinghouse. CEPI provides data to its largest user base, the public, by publishing information and data visualizations to the MI School Data website. CEPI's data also are used by state agencies, the legislature and governor's office, and researchers within and outside Michigan.

### *Approach and considerations*

CEPI stores historical data in two ways: as aggregated data in the public-facing MI School Data Portal and, in some instances, as individual-level student records that are accessed through secure requests. Individual-level data are stored primarily in snowflake schema, where a centralized fact table is connected to multiple dimensions in dimension and fact tables. As CEPI moves data to its new, standards-based data warehouse, it must determine how to store historical aggregates in the new data model using the CEDS-based data elements and dimension tables.

Stakeholder need is a major consideration for CEPI. Frequency of use, data complexity, visibility, and political climate all determine the priority of data migration. CEPI takes a multi-prong approach where a small set of complex data is paired with a broader set of more stabilized data. This approach allows CEPI to move two deliverables forward at the same time while spreading the work and associated learning opportunities among staff across multiple levels of ability. Stakeholder interest has led CEPI toward end-point development, such as migrating historical data and developing public report displays early in the overall timeline.

Limitations in technology and workload also affect data migration. Because CEPI supports daily data operations, as well as analytical needs, the storage and maintainability of the current infrastructure are scarce resources. Massive transactional logs spanning over 20 years of operation also take up key space, and resources are being shared by a parallel effort to migrate data to the cloud as a part of SLDS modernization. To make the process smoother, CEPI has started a "Direct to Data Warehouse" approach, moving data from the source data system directly to the CEDS Data Warehouse. This process should help transfer data with fewer dependencies because it allows freedom to load and leverage historical data that are not as highly structured as modern datasets.

Finally, data quality is a consideration for data migration. Rigorous data quality checks occur during the transition of data from staging to the integrated data store (IDS) and the data warehouse. However, this process can cause issues with data that are not ready to undergo these checks. For example, transactional date-based data and CEDS data are tied together based on the individual's role in the education system (such as K12 student or educator), which can cause issues during migration if the destination system is not structured similarly. Dates cannot always be created, and combining multiple datasets into a single person's role in the education system is not always cohesive. Migration issues frequently become validation issues.

### *Governance*

Data governance plays a key role in CEPI's historical data migration process. CEPI uses the governance process to finalize decisions and eliminate gaps during mapping and alignment. It also can create buy-in and support from multiple user groups. The amount of governance required for a single domain or content area of data determines its migration priority; areas with significant gaps may be delayed or have an extended timeframe planned. The collaborative nature of data governance provides added benefits around socializing ideas, creating buy-in among a broad team, and using governance as a training platform.

### *Challenges*

Several small challenges affect CEPI's data migration process. The SLDS has several dozen data systems that feed into it, along with more than 100 reporting views and nearly 800 reports that aggregate data. CEPI frequently has to determine from which level to migrate data. Individual-level data must migrate as they are the core of the data warehouse; however, aggregate data also must migrate because they have been published already and the risk of recalculating those aggregates differently is significant.

Much of the SLDS data are housed using "point-in-time snapshots" to determine historical settings instead of the effective and end dates that CEDS leverages, making the data difficult to transform. To overcome this requirement, CEPI has implemented a default June 30 and July 1 effective and end dates when the school year field is missing. CEPI also removes the date requirement from the CEDS Data Warehouse when necessary. Finally, there is not

yet a standard for aggregate storage, although CEPI is looking into developing one as a mechanism for sharing output reporting code.

## Lessons Learned

***Make the shape and format of the legacy data system as close to CEDS as possible before loading into the traditional CEDS model***

Data should have start and end date fields for transactional components and be tightly coupled. Organizations also could consider using the "Direct to Data Warehouse" approach, which effectively eliminates many data quality conundrums and requirements of a fully actualized IDS.

***Avoid burnout***

Converting historical records can be a tedious and demanding job. Have a dedicated team for the task, and switch people in and out of the team as needed to avoid burnout.

***Document your process***

Migrating historical data can be a learning curve. Be sure to document how historical tables line up with current data, as well as other key processes.

## Conclusion

As states modernize their SLDSs, they must consider what to do about the many years of data that have been collected and stored with older technologies and structures. Staff knowledge, storage, and manpower are key resources that may be affected by the move. Clear documentation and preparing historical data to match the new format before loading into the new system can help states prepare for a smooth transition.

## Additional Resources

Common Education Data Standards
*https://ceds.ed.gov/*

Michigan Center for Educational Performance and Education
*https://www.michigan.gov/cepi/*

South Carolina Department of Education
*https://ed.sc.gov/*

State of Vermont Agency of Education
*https://education.vermont.gov/*