PALDS

Pennsylvania Enterprise Research and Longitudinal Data System Project

2024 NASCIO State IT Recognition Awards

Category: Data Management, Analytics & Visualization

Project Initiation & End Dates: July 17, 2023 – October 19, 2023

State: Commonwealth of Pennsylvania

David Partsch | Chief Data Officer
PA Office of Administration | Enterprise Data Office
555 Walnut Street | 7th Floor Forum Place
Harrisburg, PA 17101

EXECUTIVE SUMMARY

The Commonwealth of Pennsylvania invests significant resources into programs that support its constituents. They include programs to improve educational outcomes, enhance workforce development, promote public safety and well-being, build thriving communities, and more. Despite having copious amounts of data on these programs, the Commonwealth has lacked the ability to look at the big picture these data can provide.

The Pennsylvania Longitudinal Data System, or PALDS, is a service of the Office of Administration (OA) that securely connects and analyzes data across programs and agencies to support improved decision-making, service delivery, program administration, and long-term outcomes. PALDS offers a unique opportunity to conduct rigorous, multi-agency research using longitudinal data. It provides research products to help Commonwealth decision-makers support the public interest and empower all Pennsylvanians to achieve better outcomes and unlock a brighter future.

PALDS is overseen by a governing board that includes OA, Governor's Office of Policy and Planning, and the Departments of Corrections, Education, Human Services, and Labor and Industry. OA contracted with Amazon Web Services (AWS) to develop an enterprise platform to securely collect, ingest, store, cleanse, master, match, aggregate, de-identify, and deliver data to analytic tools to visualize and report research findings. The 14-week implementation commenced on July 17, 2023, and completed October 19, 2023. The overall deployment cost was $819,000. PALDS is deployed in an AWS GovCloud instance and utilizes native AWS cloud services for data processing, governance, and security.

PALDS has already demonstrated its capabilities through a successful pilot project linking data across the Pennsylvania Department of Education (PDE) and Pennsylvania Department of Labor & Industry (DLI), producing actionable insights that benefit Pennsylvanians. Moreover, PALDS has increased Pennsylvania's competitiveness for external grant opportunities, resulting in an award of up to $2.9 million from the U.S. Department of Labor's Workforce Data Quality Initiative (WDQI) program.

The PALDS team is working with its partner agencies to develop a long-term, multi-agency research agenda. This agenda will be expanded over time as new research questions and use cases will be added to the PALDS project pipeline. Pennsylvania also plans to scale PALDS to an enterprise service capable of serving all state agencies as well as external stakeholders.

The Commonwealth and its agencies share a common goal of improving the lives of constituents. PALDS is bridging the data gap across agency divides to evaluate the effectiveness of our programs. This collaboration can help us identify resources that can benefit our constituents, secure funding from external partners, and provide policymakers with valuable insights to make informed decisions. Pennsylvania's investment in PALDS is an investment in its future, offering a unique opportunity to drive innovation and improve outcomes for all Pennsylvanians.

IDEA

Many public sector entities struggle with data “blind spots.” Much like the blind spots we experience while driving, data blind spots can limit our situational awareness and introduce risk into our decision-making.

Data blind spots are created by the familiar challenge of siloing—in this case the siloing of our data. When individuals move from one program or system to the next, the previous program often loses visibility into how they’re doing. Likewise, the new programs or systems lack visibility into the supports provided already to the individual. As a result, leaders and policymakers must often make decisions without fully understanding the relationships between programs and outcomes.

Data silos can be the result of needing to comply with the law or protect individual privacy, but this is not always the case. Further, data silos also place limitations on:

- Our ability to measure outcomes over time and across geography;
- Our ability to understand how co-enrollment or participation in multiple programs (e.g. reentry and workforce development) interacts to influence outcomes (e.g. employment);
- Our ability to see an individual’s history and background and tailor programming to their needs; and
- Our ability to invest in programs with a strong evidence base for meeting our policy goals.

To meet this challenge, the Pennsylvania Office of Administration (OA) created the Pennsylvania Longitudinal Data System (PALDS) to bring together education and workforce data across programs and agencies to support informed decision-making and improved service delivery, program administration, and long-term education and workforce outcomes.

A Governance Board composed of state government officials, including the Secretaries of Administration, Corrections, Education, Human Services, Labor & Industry, and Policy & Planning (Governor’s Office), was brought together to oversee the program. Through strategic planning, committees of PALDS staff and agency stakeholders have been developed for Data Governance, Research Pipeline, Education and Capacity Building, Communications, and Sustainability.

Evidence-based policymaking and investing in what works require government decision-makers to understand the outcomes associated with public programs. For example:

- How do students’ courses of study in high school affect their rate of college graduation or employment?
- Does participation in a high-quality early childhood education program promote kindergarten readiness?
- What supportive measures can be made available to help Pennsylvania residents?
- How do we most effectively invest state resources to support thriving communities?

In some cases, we can answer questions like these using a single agency’s data. Many times, however, an agency provides a service whose outcomes are measurable only by connecting its own data with data from another agency.

PALDS provides a secure environment for the harmonization of data to create statistical analyses, interactive dashboards, and other reporting tools to provide insights to make more informed decisions and improve outcomes for Pennsylvanians.

- Provide information on program outcomes
- Analyze and conduct research on programs and effectiveness
- Help direct and allocate resources to helping our constituents become successful
- Provide policymakers with information needed to facilitate decision-making

Historically, there have been legal, technical, and administrative barriers to agencies sharing data, despite the critical role data-sharing plays in evidence-based policymaking. PALDS helps to overcome these barriers by providing a program dedicated to sharing data in a manner that is safe, legal, and ethical. Data governance policies and security protections include data obfuscation and masking; data use agreements; data dictionaries; and quality assurance checks and security audits.

IMPLEMENTATION

Project Particulars: Recognizing the vital need to tie policy and program decisions to quantitative metrics and data analyses, the Commonwealth engaged in a collaborative effort to implement an enterprise research platform. A project team consisting of the Governor's Office of Policy and Planning, Office of Administration, Office of General Counsel, and agency stakeholders collaborated to design, develop, and deploy PALDS in calendar year 2023.

OA contracted with Amazon Web Services (AWS) to develop the PALDS to securely collect, ingest, store, cleanse, master, match, aggregate, de-identify, and deliver data to analytic tools to visualize and report research findings. A 14-week implementation commenced on July 17, 2023, and was completed October 19, 2023. The total PALDS deployment cost was $819,000, which included licensing, infrastructure, professional services, and personnel. PALDS is deployed in an AWS GovCloud instance and utilizes native AWS cloud services for data processing, governance, and security.

Technical Architecture: PALDS is deployed on an AWS GovCloud instance given the types of data (PII, PHI, CJIS, FERPA, etc.) being brought together across agencies and business areas for research and analytic operations. AWS GovCloud is certified at the FedRAMP high impact level. PALDS takes advantage of AWS auto-scaling, load balancing, and resiliency functions to ensure processing speed and efficiency as well as 99.999%+ uptime.

Reusable data pipelines are developed and deployed using AWS cloud services such as Transfer Family, Kinesis, Lambda & Step Functions, Glue/Data Brew, SNS, SQS, S3, DynamoDB, Athena, SageMaker, etc. to create automated processes for ingesting, storing, cleansing, mastering, matching, aggregating, de-identifying, and/or hashing data prior to making the data available to analytic tools such as Power BI, Tableau, R, SAS, etc. to visualize and report research findings. PALDS is agnostic with respect to the analytic tools used by researchers and data scientists and allows secure connection to data based on their analytical needs. PALDS is also capable of publishing data and research findings to the Commonwealth's Open Data portal (OpenDataPA) in instances where publishing to the public for transparency purposes has been approved by participating agencies.

Data Governance, Security, and Privacy: Data governance processes and technologies have been implemented to ensure appropriate security/access, privacy, and usage considerations are maintained within the PALDS.

Agencies are required to sign and participate in the Commonwealth's enterprise memorandum of understanding for data sharing. All multi-agency research projects that use PALDS require the participating agencies to enter into a formal data exchange agreement.

The Commonwealth's enterprise data catalog and data governance platforms have been integrated into PALDS to provide real-time meta-data management, data inventory, data classification and categorization, as well as data quality, data privacy, and data risk assessment and maintenance.

PALDS identity and access management/role-based access control configuration is integrated with the Commonwealth's enterprise active directory service for user authorization, access, and login control. AWS cloud services such as Secrets Manager, Parameter Store, Security HUB, CloudTrail, CloudWatch, and Lake Formation are used to centrally govern, secure, log, audit, and share sensitive system attributes, security keys, and data for analytics and machine learning functions.

For research studies where privacy of participants is required or where legal and/or regulatory compliance warrants it, data hashing with salts is utilized to match/link individual record-level data from multiple agencies without receiving PII. Once data is transferred, PALDS would match/link the data using the hashed identifiers.

IMPACT

With the implementation of PALDS, Pennsylvania gained—for the first time—a technical infrastructure and program team dedicated to conducting rigorous, multi-agency research using longitudinal data.

Following completion of the data environment deployment in October 2023, PALDS launched a pilot research project with the goal of measuring labor market outcomes of Pennsylvania's adult basic education participants. Adult basic education (ABE) programs and services support Pennsylvanians in developing basic skills in math, reading, and writing; competency in the English language; preparation for the high school equivalency test; and more. The goals of the project include providing PDE insights into the employment and wage outcomes of ABE participants over time and identifying opportunities to improve ABE program management and the delivery of adult education services.

To achieve these goals, PALDS worked with PDE and DLI to link data on ABE participants with unemployment compensation (UC) wage records in the PALDS environment. The pilot project brought together participant and wage record data from 2018 to 2023, which enabled PDE to measure participant outcomes over a longer period than the annual snapshots the agency captures for federal performance reporting. For example, the project yielded findings on outcomes including:

- The percentage of adult education participants who were employed at different intervals (e.g. one, two, and three years) following their enrollment in adult education;
- The wages that adult education participants earned at different intervals (e.g. one, two, and three years) following their enrollment in adult education;
- The relationship between the number of instructional hours participants took and the wages they earned during the same period of time;

- Variation in instructional hours and wages across demographic groups, such as those defined by gender, age, race, and ethnicity.

To report project findings, the PALDS team developed and published an interactive data dashboard that is available to PDE's Division of Adult Education. The dashboard provides program administrators with new insights into the labor market outcomes of the population they serve through data visualizations that are filterable by time period, instruction type, hours of instruction, and participant demographics. Through the pilot project, PALDS significantly enhanced the insights available to decision-makers by bringing together data over time and across agencies to measure ABE participant outcomes and visualizing the results of the data analysis in an interactive, user-friendly business intelligence tool. Similar insights and tools were not previously available from data systems that were designed principally for annual reporting and due to limited state resources dedicated to multi-agency research, data analysis, and data visualization.

Building on the success of the pilot project, PALDS is actively pursuing a second research project measuring the employment, social, and health outcomes of reentrants released from Pennsylvania state correctional institutions. This project is a collaboration with the Department of Corrections (DOC), Department of Human Services (DHS), and DLI, all of which will be able to advance research and policy analysis priorities through the project's findings. The project will provide a secure environment for bringing together reentrant data from DOC, UC wage records from DLI, and Medicaid data from DHS to enable the participating agencies to better understand, and thereby improve, reentrants' experience with Pennsylvania's reentry, workforce development, and health care services.

PALDS continues to work with its partner agencies to develop a long-term, multi-agency research agenda that previously did not exist in Pennsylvania. Future PALDS research projects originating from this agenda will address a set of research questions that state agencies have raised as high priorities for informing policy- and decision-making, including:

- How do the educational and labor market outcomes for students who completed career and technical education (CTE) coursework and programs of study in high school compare with students who did not complete CTE coursework in high school?
- How does participation in a pre-apprenticeship program impact an individual's credential attainment and/or employment outcomes? How does the impact vary by demographic subgroup?
- How do children who participated in high-quality childcare programs perform academically relative to peers who did not participate in these programs?
- How can DHS serve young people with disabilities and children involved with child welfare more effectively as they age out of the K-12 system?

The research agenda development process is ongoing, meaning that new research questions and use cases will be added to the PALDS project pipeline over time. In the near term, the research priorities of the PALDS Governance Board agencies will take priority on the agenda. However, in the medium-to-long term, PALDS intends to scale to an enterprise research service capable of meeting the research design, data analysis, and reporting needs of any Commonwealth agency and select external partners. In an example of the demand for PALDS's services, the Pennsylvania Workforce Development Board adopted a recommendation in February 2024 calling for PALDS to undertake a research study on the employment and education outcomes of Pennsylvanians who have participated in youth programming. PALDS is currently working with the Workforce Development Board to scope a project that responds to the Board's recommendation.

Beyond providing insights to inform policy and program management decisions, the impact of PALDS extends to increasing Pennsylvania's competitiveness in securing grant opportunities. The Workforce Data Quality Initiative grant program, administered by U.S. Department of Labor (DOL), provides funding to support state governments in developing and expanding their data systems to measure and analyze the performance of education and workforce training programs. In July 2023, Pennsylvania was awarded a WDQI Round 9 grant through DLI to fund a project that will task PALDS with deploying an advanced identity resolution and universal identifier software solution. This tool will facilitate data linkage across state agency systems that lack a common identifier, such as the Commonwealth's K-12 and workforce systems. As a result of the successful partnership between DLI and PALDS, Pennsylvania will receive up to $2.9 million over three years to strengthen the state's capacity to measure education and employment outcomes over time, evaluate the effectiveness of employment and training programs, and improve the quality and responsiveness of Pennsylvania's workforce development system.

The impact that PALDS has achieved has come through an efficient use of resources. Initial deployment and support costs for the enterprise research platform were significantly lower compared to other states. The Commonwealth spent a total of $819K on AWS infrastructure and professional services to implement the PALDS. By comparison, there are public references to the initial funding levels states such as Maryland ($2.5M), Minnesota ($5M), Utah ($7.1M), California ($3.26M), North Carolina ($3.64M), and Virginia ($7.5M over 2010-13) expended to deploy their longitudinal data systems. Combining the PALDS's cost to implement with the fact it can be used for any enterprise research project—not just education and workforce research—clearly shows a significant return on investment. [1]

Time to market for completing research studies has been reduced with the utilization of AWS cloud and DevOps services, which have allowed the Commonwealth to build reusable data pipelines to automate nearly 100% of the data process flow. Data availability and reuse has been increased as the PALDS provides a secure data marketplace for aggregated and de-identified datasets that can be requested and used by other entities/parties through appropriate governance and approval processes.

Looking ahead, PALDS will continue to carry out projects that are responsive to its partner agencies' needs for policy- and operations-relevant research and data analysis. The program will develop a research agenda that continuously evolves to meet the needs of a broadening group of stakeholders, eventually reaching the scale of an enterprise research service that is also equipped to serve external audiences through its research and data products. To support this long-term plan, Pennsylvania is providing funding for PALDS through the Office of Administration, which houses the program. Additionally, Pennsylvania will continue to receive support for PALDS from U.S. DOL for the balance of the WDQI Round 9 grant. Moving forward, PALDS will help make Pennsylvania a competitive applicant for other grant opportunities, owing to the program's ability to measure, monitor, and evaluate the effectiveness of grant dollars that flow to the Commonwealth.

[1] https://nationalskillscoalition.org/wp-content/uploads/2020/12/Cost-of-State-Longitudinal-Data-Systems_web.pdf; https://www.nascio.org/wp-content/uploads/2020/09/2018-MN-CBCP-Early-Childhood-Longitudinal-Data-System.pdf; http://slds.rhaskell.org/state-profiles