



Early Childhood Participation
Building Data Pipelines
to a more Resilient Tomorrow

Data Management, Analytics & Visualization

State of Illinois
Dessa Gypalo
Chief Data Officer
Dessa.Gypalo@illinois.gov

Initiation Date: December 2022
End Date: March 2023

Executive Summary

The Illinois Longitudinal Data System (ILDS) is a collaborative body of state agencies whose goal is to empower leaders in the education space with actionable information to support equitable outcomes and engaged communities. In short, its purpose is to ensure that high quality data about learners from birth to career will be readily available for appropriate analysis by state government program officials. In 2022, the ILDS executed data sharing and access agreements across participating agencies to ensure data flow, established data analytics priorities for the program, and launched the first of its kind data pipeline project.

The Early Childhood Participation Data Set (ECPDS) project leveraged the data provided by the ILDS agencies. This project brought together child-level data from all early childhood programs in Illinois, including programs such as the “Child Care Assistance Program” at the Illinois Department of Human Services (IDHS) and “Preschool for All” from the Illinois State Board of Education (ISBE). The ILDS program and data architecture provided the perfect tool for this effort since it incorporates data from both agencies.

To implement this project, the ILDS program team worked with both DHS and ISBE to establish “data pipelines” with the help of the Illinois Department of Innovation & Technology (DoIT), IBM, and two university partners. To enable the project, each state agency first had to stage their data. Once the data was appropriately staged, the data was virtualized into the evaluation environment. This provided the agencies with significant control over whether their data can be used as part of the ILDS data architecture.

Using a Master Client Index or key that was previously established in 2016, the data was matched across the two agencies. The value of the key is that neither agency had to provide demographic data or Personally Identifiable Information (PII) to the ILDS. Rather, all matching can be done in the CloudPak4Data (CloudPak) environment without using sensitive data. After establishing the data virtualization, a project was established that allowed for all the data to be linked. This data set is used to support research, the development of analysis, and the development of tools both for internal and external use by agencies and, when appropriate, public audiences. Development, current and future, will be done in a highly secure manner by authorized users that have been provided rights using a templated data agreement that all the agencies have agreed to.

Future work is to determine analyses that are needed by agencies and the public and to develop the appropriate tools using the developed data set.

Idea

The State of Illinois intends to be the best state in the country to raise a young child. To create this environment, Illinois must better understand the provision of early childhood services to children from birth to kindergarten and then in the initial years of formal education. Working with IBM, DoIT initiated two Design Thinking workshops to unpack this challenge. The initial workshop focused on what would be necessary to bring together data from multiple agencies that provide services for early childhood. This work highlighted the need to provide a unified data request process that could leverage a common platform to streamline requests.

The second workshop focused on answering the question “how many students receive a continuum of early childhood services?”. It was determined that an answer would require bringing together data from IDHS and ISBE. In addition, a new data set was required to better understand Head Start participation

since this program represents about one-third of all early childhood spending in Illinois but is provided directly from the federal government to grantees.

The State of Illinois provides early childhood services to approximately 150,000 children. Despite the volume of services provided, it has been difficult to count how many students receive a continuum of early childhood services – starting with home visits through kindergarten. While attempts had been made in the past to better understand the provision of services, a longitudinal data set that could be used to support a variety of data requests and tool development had never been developed for this population. This challenge initiated the ECPDS.

With the advent of the Preschool Development Grant Birth through Five (PDG B-5), the decision was made to implement the ECPDS project within the ILDS data architecture. This was done while re-envisioning the ILDS to be a centralized hub for education and workforce data. This new version of the ILDS leverages CloudPak, an enterprise data management and analytics solution supported by DoIT.

The development of the ECPDS project did not call for a traditional data warehouse to store and manage data but did require an environment to catalogue data and virtualize data sets that could eventually be combined for analytics and research purposes. The DoIT-managed environment allowed for data to be combined for a variety of purposes without ever being stored in one centralized location. This gave state agencies ultimate control over their data – by allowing agencies to share easily without shipping data files to external partners – while maintaining the ability to revoke access to the data at any time.

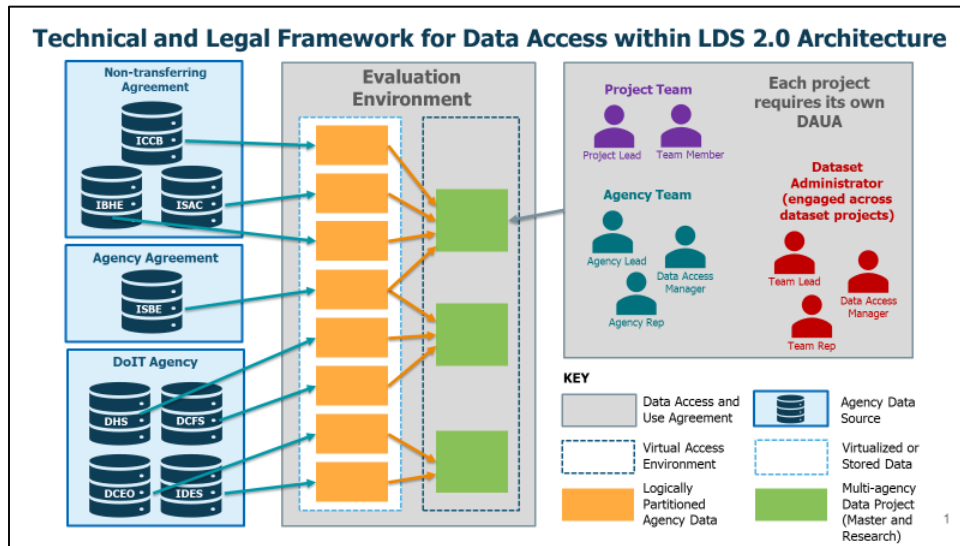
The ECPDS solution served as a pilot project for a universal challenge. Many states are working to better understand their early childhood participation while developing streamlined data sharing solutions that allow data owners to maintain control over how and with whom data is shared. This project aligns with efforts by the federal government to support states' abilities to count the number of participants accurately and precisely in early childhood programming through the PDG B-5 grant. This project also supports the State of Illinois' efforts to stand up "Birth to Five" councils which focus on improved allocation of early childhood funds at the local level. The successful implementation of this project will result in dynamic, interactive reporting around capacity, demand, and participation that can be used by local communities to better understand the impact of early childhood programming.

Implementation

The roadmap for this project was to increase the usage of the CloudPak system and connect the ECPDS to the environment. To start the project, the program lead re-engaged the ILDS governance council and established a new governance agreement. This agreement allowed for the creation of additional governance committees including the Executive Committee, which includes the eight executives of all ILDS agencies, and the Managing Committee, which includes technical and data leadership for each engaged agency. The Executive Committee meets twice annually, and the Managing Committee meets four to six times per year. Managing Committee members participate in the Executive Committee meetings.

This governance body approved both the ECPDS project and agreed to the use of CloudPak as a centralized environment for analytics projects. This will allow future data sets to be connected to the ECPDS. For example, the Illinois Department of Children and Family Services is working with the ILDS program team to connect a Children in Care data set to the ECPDS to better understand the early childhood services being provided to children in foster care.

This project is intended to serve as a pilot of an approach to data management and analytics at the domain and enterprise levels. To support the needs of domain level analytics, the ILDS program team created data governance processes across agencies starting with data inventory through connecting data to the data hub. Additionally, processes developed around data matching and data virtualization can be scaled and adapted to meet the needs of program specific analytics projects through domain and enterprise level analytics projects.

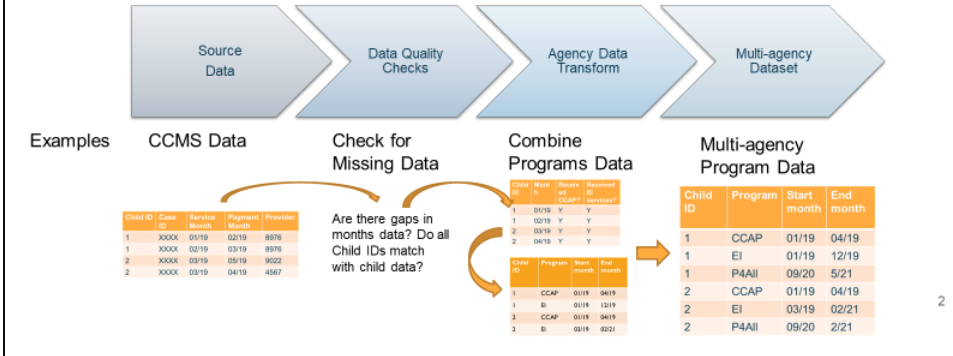


The second part of the roadmap focused on connecting data from the individual agencies and matching it in a multi-agency project. This required a data sharing agreement between two agencies and an evaluator. Each participating agency and the evaluator provided the names of team members that would participate in the project and have access to data. Each team member needed an account in the CloudPak system. In addition, each agency was required to choose a data steward – an individual or team that understood the agency’s data and can manage access to the data. This “self-service” approach to data management and access means that DoIT can support the underlying system without having access to data.

Once the data agreement was in place and the data stewards identified, the ILDS program team worked with each agency to understand how they would make their data accessible and ingest it to the CloudPak system. Working with IBM, the team initiated two data pipeline workstreams. The first workstream entailed pre-processing IDHS and ISBE data so that it was prepared to be shared in a common project. The data was then connected to CloudPak and an ILDS project using data virtualization. The second workstream involved cataloging data once it was connected to the system.

Developing Data Pipelines to Transform Data

CloudPak provides tools that allow the quality of data to be checked and allow the data to be transformed for analytic purposes:



The key output of the project was the development of a data set that can be used for a variety of purposes and research questions. To date, the teams involved have demonstrated that the data set can be used for analytics – an end state that was previously elusive. These analytics allow for a better understanding of child pathways through the early childhood system. While data will continue to be integrated into the data set (including Head Start data), initial analysis has already shown the value and long-term potential of the system. With the development of the pipelines completed, work is starting on the creation of dashboards for insights to be used at the state and local levels.

Another key criterion for this project was to increase the level of collaboration between agencies, “portal partners,” and researchers. Illinois has had a long history of outsourcing data work. While the “portal partners” that have done this work have strived to engage with the community and make sure that they create windows on agency data, the use of “portal partners” has often caused a distance between the agency and users of data including local groups and advocates. With the creation of a shared environment for data analytics, now agencies, portal partners, and researchers can more closely collaborate. For example, this project has already been leveraged by a joint team of researchers from the University of Illinois, the University of Chicago, and Northern Illinois University on another project.

To drive iterative development, the project team met weekly to review tasks lists, identify dependencies, and share learnings from similar projects. This high-touch stakeholder and team member engagement resulted in an initial version of the ECPDS in early 2023 despite delays with data access and signing of legal agreements.

The governance agreement also envisioned the creation of a steering committee to support future projects. In the case of the ECPDS, the agencies, researchers, and developers met monthly to review the requirements of the project and get updates on successes and challenges. Each member provided a unique perspective on how to bring to data together, how to use data, and how to appropriately share data in ways that are secure and maintain privacy, while supporting its use for a variety of purposes. Committee members also support ongoing communications with their respective constituencies, keeping the early childhood and larger education and workforce community engaged with progress. This level of collaboration would not be possible without a strong executive sponsor, the resources provided agencies to support data sharing, and the work of DoIT.

The underlying technical architecture built is not only intended to support the ECPDS but scaled to support data analytics and management across the State of Illinois. Elements of this architecture include

the use of a centralized data management and analytics system behind DoIT's firewall that means that data is not shared externally. The environment supporting the project is within DoIT's data center, which requires adherence to all the state's security and hosting requirements. Additionally, the architecture approach of the environment minimizes the amount of data storage within the analytic environment and avoids a single point of failure for access to the data.

As part of this project, processes for data sharing and de-identifying data were developed to ensure privacy. Approaches to data use outlined in data agreements and the requirements to sign security pledges mean that both organizations and individuals understand the sensitivity of the data they are working with and appropriate uses of that data. Development of the ECPDS also required data governance processes to be defined and implemented, such as a process for de-identifying and matching disparate data sets to ensure privacy of individuals' data.

Impact

The impact of this project goes far beyond the analytic results of the creation of the ECPDS. In addition to creating a data set that increases understanding of early childhood services across the State of Illinois, this project resulted in a proof of concept for how the collaborative approach to data sharing and data governance can be scaled across projects and domains.

Prior to this pipeline development project, data sharing for early childhood data required identifying appropriate data owners and data stewards, which was an arduous process for requesting appropriate data fields and formats and was for single use only. This project took a new approach to data sharing and data set creation by virtualizing data, allowing data to be used and reused for multiple projects, and creating a collaborative in-house environment for analysts and researchers to access data more securely. Additionally, the approach used for this project allows for research and analytics findings to be replicated and expanded for future projects.

While it is not yet possible to evaluate the return on investment of this project, it is worth noting that the State of Illinois spends over \$1.5B dollars annually for early childhood services and the expectation is that this will grow over time. If this project leads to 1% better allocation of funds, it will result in over \$15M in improved investment. The program team estimates that ongoing support for this project will be less than \$1M per year once all the processes have been automated, with some of this cost shared between data projects (for example, the data stewardship of data at ISBE). Agencies are already envisioning how insights generated by this effort could be used to revamp programming.

Feedback from the "portal partners" has demonstrated the value of the collaboration and work has begun on report design that can leverage the data being provided. Participating agencies will also have better analytical capabilities for their own programs and the ability to better understand their programs interactions with other programs with this collaborative approach to research and analysis.

With this project completed, the next steps for this project include the development of front-end reporting and visualization of this data that can be used at the local community level. This work includes disaggregating the early childhood data by region and demographics and providing analysis of the needs of regions, their enrollment, their capacity, and levels of funding. Each of these analyses will be layered across programs to ensure participation can be understood and measured across the early childhood continuum. The goal is to eventually expand this work into other Health and Human Services including Medicaid, SNAP, and TANF to provide the same level of research and analysis across the State of Illinois domain programs.