



Building DC's First P20W+ Longitudinal Data System

Architecture, identity resolution, and governance for a cloud-native education-to-workforce data system on Azure Databricks

CONTENTS

02

DC'S VISION FOR A CONNECTED DATA SYSTEM

03

UNIFYING DATA THROUGH IDENTITY RESOLUTION USING P3RL METHODOLOGY

04

PLATFORM ARCHITECTURE

06

DATA INTEGRATION AT SCALE

06

INITIAL RESEARCH OUTPUTS

07

KNOWLEDGE TRANSFER AND DISTRICT OWNERSHIP

07

LESSONS FOR OTHER JURISDICTIONS

INTRODUCTION

Building a P20W+ longitudinal data system requires solving problems and navigating data engineering, privacy, governance, and multi-agency relationships simultaneously. This paper documents how the [District of Columbia's Office of Education Through Employment Pathways \(ETEP\)](#), in partnership with Resultant, designed and implemented one of the first P20W+ systems built on [Azure Databricks](#), and what that process revealed about doing this kind of work well.

The effort was supported by a strong governance framework that included ETEP's data governance and research committees, which help guide data-sharing decisions, research priorities, and responsible use of integrated data. The project also benefited from close collaboration with the District's Office of the Chief Technology Officer (OCTO), which serves as a key infrastructure partner supporting the District's enterprise data environment.

DC'S VISION FOR A CONNECTED DATA SYSTEM

ETEP, within the Office of the Deputy Mayor for Education, is responsible for understanding how DC residents move through its public education and workforce systems. That requires connecting records over time through early learning, K-12, postsecondary programs, workforce training, and wage records. Those records existed across multiple partner agencies with no unified way to link them. The challenge was to design a sustainable system that could connect these datasets with a common identity layer, governed integration, and infrastructure built for longitudinal analysis at scale, without exposing personally identifiable information (PII) to researchers or downstream systems.

From the outset, ETEP and Resultant designed and implemented the system to comply with applicable federal and District privacy laws, regulations, and security requirements. This included adherence to the Family Educational Rights and Privacy Act (FERPA), agency-specific Data Sharing Agreements (DSAs), District cybersecurity standards, and established policies governing the collection, storage, transmission, and use of sensitive data. Privacy, security, and responsible data stewardship were treated as foundational design requirements rather than technical considerations addressed later in implementation.



UNIFYING DATA THROUGH IDENTITY RESOLUTION USING P3RL METHODOLOGY

Even well-governed data tells only a partial story without a reliable, secure way to link individuals across systems. Solving this reliably and at scale was just one of several consequential technical challenges in the project.

WHAT P3RL IS

P3RL (Privacy Preserving Probabilistic Record Linkage) is Resultant's proprietary identity resolution solution. It assigns a unique identifier to each individual across all contributing datasets, making longitudinal analysis possible while keeping raw PII secure throughout the process.

What makes P3RL distinctive is its ability to resolve identity across datasets that were never designed to share a common identifier, and to do so reliably at scale. Real-world government data carries the kinds of quality issues that rule-based matching can't handle without building an ever-growing and increasingly fragile ruleset.

HOW IT WORKS

P3RL draws on multiple layers of matching logic, including:



Blocking to narrow candidate record pools before matching begins, limiting computational overhead



Rule-based fuzzy similarity scoring across multiple PII fields with configurable weights



A supervised machine learning layer that refines match confidence based on patterns in the data



Human-in-the-loop review for ambiguous cases that fall below confidence thresholds

Domain knowledge rules are encoded as version-controlled configurations, keeping matching logic transparent, auditable, and adaptable as the system grows to include new data sources. Each new source added to the system extends coverage and strengthens the accuracy of matches already made.

The result is a unique identifier assigned to each individual across all data-contributing agencies, creating the common thread that makes longitudinal analysis possible.



PLATFORM ARCHITECTURE

AZURE DATABRICKS FOR SECURITY AND SCALE

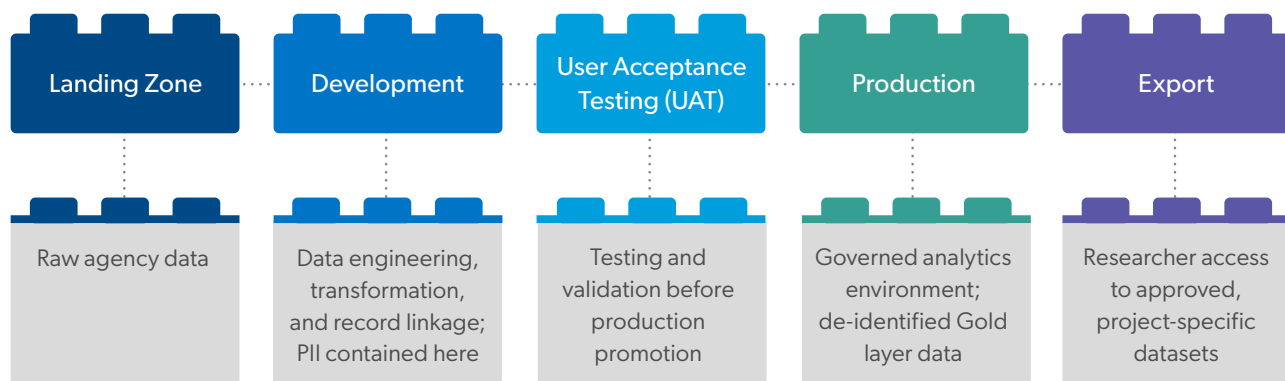
ETEP prioritized building the system in a way that maximized existing DC infrastructure and investments. The resulting cloud-native architecture, built on Azure Databricks, aligned with the District's technology strategy while providing the security, governance, and scalability needed to support the P20W+ system as it grew.

Azure Databricks serves as the foundation for the full data lifecycle: acquisition, ingestion, transformation, record linkage, governed access, and researcher access to approved datasets. Consolidating those capabilities within a single environment reduces integration complexity, simplifies governance, and makes the system easier for a lean team like ETEP to own and evolve.

The architecture was intentionally designed to support the District's privacy and security obligations. Data segregation, controlled movement of data between environments, role-based access controls, encryption, audit logging, and de-identification processes were implemented to align with federal privacy requirements and District security policies. These controls help ensure that sensitive information remains protected throughout the data lifecycle while still enabling approved research and longitudinal analysis.

WORKSPACE STRUCTURE

The system is structured as five purpose-built Databricks workspaces, each with a distinct role in the data lifecycle and its own dedicated storage container:



Data moves between workspaces only through controlled, orchestrated mechanisms. PII is strictly contained in the first two workspaces. By the time researchers access data, it has been cleaned, linked, de-identified, tested, and governed.

MEDALLION ARCHITECTURE

Within the Development workspace, a medallion architecture organizes data through three progressive quality layers before curated datasets are promoted downstream:



BRONZE: raw ingested data



SILVER: cleaned and standardized data with cross-agency identity linkage



GOLD: de-identified, aggregated, research-ready datasets; the only layer researchers access



This structure ensures data quality is incremental and auditable. Each layer's transformations are documented and version controlled.

GOVERNANCE VIA UNITY CATALOG

Unity Catalog provides centralized governance across all workspaces, enforcing permissions at the catalog, schema, and table level. Every query a researcher runs is executed under their own security context; they can only see data they are individually authorized to access.

Audit logs, data lineage tracking, and version-controlled configurations give ETEP full visibility into how data moves and changes across the system. This is a critical requirement for a research environment; every finding needs to be traceable back to its source.



DATA INTEGRATION AT SCALE

In its first two years, the Education Through Employment (ETE) Data System has ingested 30 datasets from six partner agencies, covering early learning, K-12 enrollment, K-12 program participation, postsecondary enrollment, workforce training, social service program participation, and wage records. The system enables longitudinal analysis of more than a decade of education and workforce history.

Each dataset required agency-specific ingestion logic, field mapping, and standardization before it could enter the pipeline. Inter-agency DSAs governed what data could be shared and under what conditions.

ETEP led the inter-agency DSAs across partner agencies, a significant coordination effort that ran parallel to technical development. Resultant provided advisory support and documentation to help the team prepare.

In addition to establishing DSAs, ETEP's governance and research committees provide ongoing oversight for data access, project approvals, and research priorities. These structures help ensure that data use remains aligned with agency objectives, community interests, and established privacy protections.



INITIAL RESEARCH OUTPUTS

The system's initial research outputs are a set of Tableau dashboards planned for public release later in 2026, covering:



**MEDIAN WAGE
TRAJECTORIES**
BY EDUCATION PATHWAY



**TOP INDUSTRIES
OF EMPLOYMENT**
BY ATTAINMENT LEVEL



**WAGE
OUTCOMES**
BY INDUSTRY

These dashboards will give DC policymakers their first comprehensive, governed view connecting education program participation to wage and employment outcomes for DC public and public charter school students, not just those required for federal reporting. Informed by extensive engagement with agency partners and public stakeholders, they serve as a starting point for exploring critical questions about the impact of education on economic opportunity. As the research agenda grows, the system is designed to take on new policy questions and add new data sources, giving leaders an expanding foundation in alignment with evolving needs.

KNOWLEDGE TRANSFER AND DISTRICT OWNERSHIP

From the start, the engagement was designed so that ETEP would fully own what was built. Resultant embedded training and documentation throughout delivery rather than scheduling knowledge transfer as a final handoff, and design decisions were explained as they were made so the team could understand not just how the system works, but why it was built the way it was.

Resultant worked as a thought partner alongside the ETEP team throughout implementation. That meant ongoing collaboration on technical decisions, data governance practices, and how to extend the system as the research agenda grows, building institutional knowledge in parallel with the system itself.

"The Education Through Employment Data System was designed with a focus on data privacy and data governance, and to be responsive to the questions and priorities of both agency partners and the community. Paired with the technical expertise of the Resultant team, we were proud to deliver a system that meets the District's current needs and can continue to evolve,"

MONICA DODGE,

Executive Director, DC Office of Education Through Employment Pathways



LESSONS FOR OTHER JURISDICTIONS

DC's ETE Data System is one example of what becomes possible when education and workforce data are connected across agencies. Many states and municipalities have the underlying data but haven't yet operationalized it into a system that reveals how education investments translate into workforce outcomes. A few observations from this project that transfer:



Identity resolution is one of the most consequential problems to solve in a P20W+ system. Without a reliable, privacy-preserving way to link individuals across agency systems, each agency's data can only tell its own part of the story. Getting this right early is what makes cross-agency longitudinal analysis possible at all.



Platform selection should follow client needs, not the other way around. For DC, that process led to building on Azure Databricks, a platform already in use across several agencies. Starting from a clear understanding of what ETEP needed, who would use the system, and how it would grow shaped every technical decision that followed.



Governance needs to be configured into the system from the beginning. On this project, that meant building Unity Catalog's access controls, audit logging, and data lineage tracking into the architecture as it was designed, not after the fact. A research environment that can't demonstrate how data moves and who can see it won't earn the trust of partner agencies or researchers.



Inter-agency data sharing requires dedicated project management, not just technical coordination. ETEP led the DSAs across partner agencies. Keeping that workstream moving in parallel with technical development was essential to keeping the project on track.



On this project, embedding knowledge transfer throughout delivery rather than scheduling it at the end made a measurable difference. Design decisions were documented and explained as they were made. The goal from the start was for ETEP to understand the system fully, not just operate it.

ABOUT RESULTANT

Resultant works with state and municipal agencies to design and implement P20W+ data systems built for long-term use. Connect with our Education and Workforce team at resultant.com.

