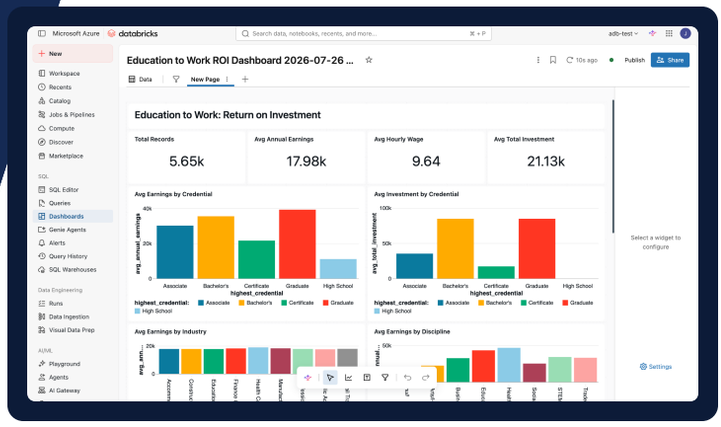
DCP

Paxcel Technologies
DATA COLLABORATION PLATFORM

Higher Education's Return on Investment in a Multi-State Region



How the Data Collaboration Platform replaces manual data prep with reusable templates and automated, privacy-safe pipelines — connecting results across agencies & states.

What is the Data Collaboration Platform?

To measure the ROI of a college credential, an analytics team needs two very different kinds of data side by side: what people studied, and what they went on to earn. That data lives in separate systems, is owned by different states & agencies, and arrives in a different shape every time — carrying SSNs and other sensitive identifiers throughout.

The **Data Collaboration Platform** turns that recurring, manual effort into a one-time setup. Cleansing, validation, anonymisation, and matching rules are defined once as reusable templates, assembled into a flow, and run automatically on every future data drop — feeding a live dashboard with no hand-off in between.

Platform Capabilities

Data Anonymiser

Protects SSNs & PII automatically, so raw identifiers never flow downstream.

Data Cleansing

Standardises messy, multi-source inputs into one consistent shape.

Data Validation

Catches bad or non-conforming records before they spread into results.

Data Matching

Links education and workforce records for the same person, across sources.

~1.2M
individual-level records processed across the pipeline
DCP CASE STUDY

5
U.S. states contributing education & workforce data
DCP CASE STUDY

5
automated capabilities chained into every flow
DCP CASE STUDY

100%
of pipeline steps logged & auditable end-to-end
DCP CASE STUDY

"The recurring pain was never doing this once — it was doing it again for every new source and every refresh." Templates and flows remove that repetition permanently, replacing it with a pipeline that runs itself.

Our Solution & Methodology

Each capability is configured once as a template and saved to a library, then reused on every source and every run. Templates are dropped into a flow in the order they should execute: the education and workforce flows chain Cleansing → Validation → Anonymiser → Collection, while a third flow matches the two sides and collects the linked result. Once defined, a flow runs as an automated, fully-logged pipeline — re-running for a new data drop needs no rebuild.

Data Volume Processed

Individual-level records by type

Education records

~1,005,000

Workforce (wage) records

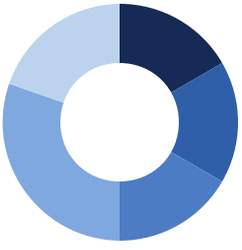
~200,000

Total processed

~1,205,000 records

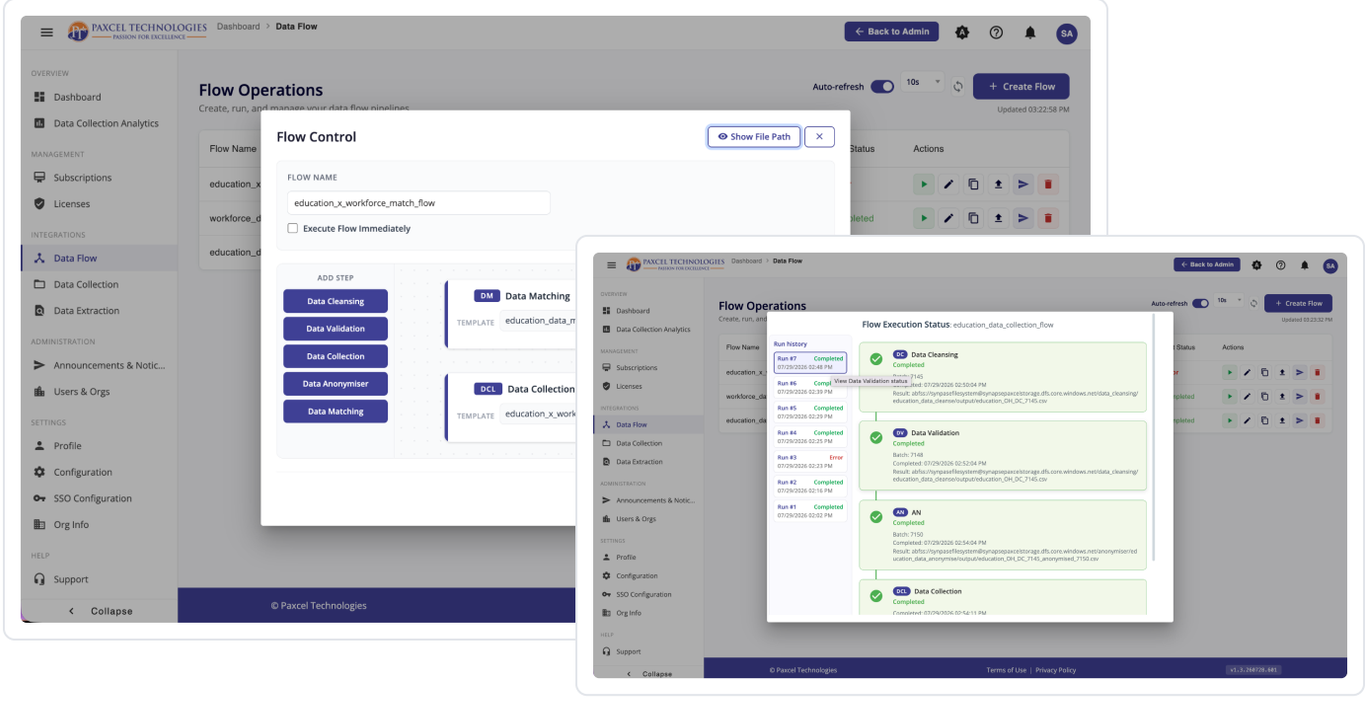
Capability Usage Across the Pipeline

How often each capability fires per full run



- Collection — 3 runs
- Cleansing — 2 runs
- Validation — 2 runs
- Anonymiser — 2 runs
- Matching — 1 run

Tools



Flows are assembled from templates, then every run is logged step-by-step in an auditable execution history.

The Outcome

The platform took ~1.2M records from five states to a working ROI dashboard, with sensitive data protected throughout.

01 **Set up once, not every time**

Templates and flows are defined a single time, then reused across all five sources and every refresh.

02 **Privacy by default**

Anonymisation runs as a pipeline step on every execution, not dependent on someone remembering to do it.

03 **Repeatable matching at scale**

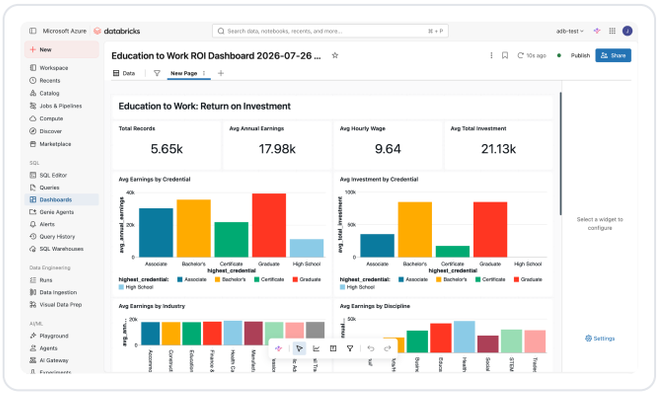
The two collections linked cleanly, with a predictable, reviewable match rate across a million-plus records.

04 **Always-current insight**

Collection feeds Databricks directly, so the dashboard reflects the latest run without a manual export.

05 **Auditable by design**

Every flow logs each step, so results can be traced back to how they were produced.



The finished ROI dashboard — always current, since it is fed directly by the pipeline.

0

manual exports between the pipeline and this dashboard

The net effect: the time, cost, and compliance risk that used to accumulate with every new data drop are replaced by a pipeline that runs itself.

Ready to see it on your own data?

The same templates, flows, and pipelines shown here can be configured for any multi-source, multi-state data collaboration use case.

About this case study

This is an illustrative use case built to demonstrate the capabilities of the Data Collaboration Platform. It does not represent a specific named organisation. The datasets are synthetic and all identifiers are fabricated; figures describe this demonstration and are not a client's reported results.